TESIS DOCTORAL

Sistema de Interacción Humano-Robot basado en Diálogos Multimodales y Adaptables

(Human-Robot Interaction System based on Multimodal and Adaptive Dialogs)

Autor:
Fernando Alonso Martín
Directores:
Miguel Ángel Salichs Sánchez-Caballero
Javier Fernandez de Gorostiza Luengo

DEPARTAMENTO DE INGENIERÍA DE SISTEMAS Y AUTOMÁTICA

Leganés, 13 Octubre 2014

TESIS DOCTORAL (DOCTORAL THESIS)

Sistema de Interacción Humano-Robot basado en Diálogos Multimodales y Adaptables (Human-Robot Interaction System based on Multimodal and Adaptive Dialogs)

Autor (Candidate): Fernando Alonso Martín

Director (Adviser): Miguel Ángel Salichs Sánchez-Caballero
CoDirector (Adviser): Javier Fernandez de Gorostiza Luengo

Tribunal (Review Committee)

Firma (Signature)

Presidente (Chair): _____

Vocal (Member): _____

Secretario (Secretary): _____

Título (Degree): Doctorado en Ingeniería Eléctrica, Electrónica y Automática
Calificación (Grade): _____

Leganés, de de

Al recuerdo de Luis Alonso Busnadiego,
mi padre, que estará siempre presente.

"Si se puede imaginar... se puede programar."
— Fernando Alonso Martín

Agradecimientos

Es difícil entender la importancia de los agradecimientos de una tesis doctoral hasta que no se ha terminado. En ese momento te das cuenta de cuánto tienes que agradecer a tanta gente. Intentaré resumir en unas líneas la gratitud que siento a todas las personas que han estado presentes durante esa etapa, haciendo posible que hoy deje de ser un sueño para ser una realidad.

Al Dr. Miguel Ángel Salichs, director de esta tesis, por toda su ayuda y sabios consejos. Le agradezco que me haya abierto hace ya varios años las puertas del grupo de trabajo RoboticsLab, dándome la oportunidad de tener una visión más amplia del mundo de la investigación y descubrir cuánto me motiva. Agradezco también al Prof. Javier Gorostiza, el codirector de la tesis, por su disponibilidad y colaboración. Especialmente el trato cercano y cordial que hemos mantenido y espero conservar.

A María Malfaz, que tanto me ha ayudado en la elaboración, redacción y escritura de artículos, tesis de master y este documento. A Ramón Barber, por ser una persona tan entrañable y predispuesta a ayudar siempre en la medida de lo posible. A Álvaro Castro, siempre con una sonrisa en la boca y dispuesto a trabajar en grupo, y por su puesto a su paisano Alberto Ramón, al que los mismos calificativos se le pueden aplicar. Al monsieur Arnaud Ramey, que vino desde Francia para revolucionar el grupo y hacernos avanzar a velocidad de crucero. A Victor Gonzales-Pacheco y su visión global de la tecnología y del software. A David G. Godoy por ponerme a mi disposición el hardware adecuado en cada momento, que desde el comienzo de esta tesis ha sido mucho. A los nuevos compañeros Jonathan y Ester con los que ya hemos vivido momentos entrañables, y especialmente a Irene, que se ha tomado la molestia de revisar en profundidad la redacción de este documento. No me gustaría olvidarme de los que ya no están, como fueron los técnicos de laboratorio Quique y Luís; Ana Corrales, que fue la primera persona del grupo que conocí y que también

fue mi profesora de prácticas de robótica en mi etapa como alumno. A los innumerables proyectantes que han pasado por el laboratorio durante estos años, que también pusieron su granito de arena, como fueron Enrique, Genis, Oscar, Javi, etc. No me gustaría olvidarme de los miembros del departamento, que pese a no formar parte del grupo de trabajo propiamente dicho, también han aportado ideas, traducciones o simplemente su amistad.

A los supervisores de las estancias internacionales, el profesor Joao Sequeira del IST (Lisboa), y a Tony Belpaeme de la Universidad de Plymouth (UK), por su agradable recibimiento y colaboración. Especial dedicación a los compañeros y amigos que conocí en esos tres meses, en especial a Jorge Gallego Pérez (Madrid), Jacobo Valle (Galicia), Tiger (Taiwan), Aitor Álvarez (Asturias), Jimmy y Edwin (Ecuador)... gente de muy diversos lugares que me hicieron sentir como en casa.

Al gobierno de España y a la Unión Europea, que han financiado varios proyectos en los que se ha enmarcado esta tesis: "Peer to Peer Robot-Human Interaction", del MEC (Ministerio de Ciencia y Educación), "A new approach to social robotics", del MICINN (Ministerio de Ciencia e Innovación). Sin esta financiación, especialmente en estos años de crisis económica, no hubiera sido posible el desarrollo de este trabajo. Espero poder devolver la inversión realizada en mi persona, con la presente contribución a la comunidad científica.

A mis padres por haberme enseñado que la vida es para los valientes. A mi hermana Ana Belén que siempre está ahí para lo que sea y a mi hermano mayor Luis María, por marcar el camino de estudio que posteriormente seguimos los demás hermanos. A mi madre Leonor, que además de darme la vida ha estado siempre pendiente de mis luchas diarias. A mi padre Luis, que pese a que ya no está con nosotros, le debo agradecer gran parte de lo que soy. Espero que si no nos hubiera dejado, estaría orgullo de mi.

A mis amigos, que han sabido disculpar mis ausencias y siempre han tenido una palabra de ánimo. Estoy absolutamente convencido de que si de algo puedo presumir en esta vida es de los grandes amigos que tengo, lo que me hace sentir una persona muy afortunada. No hace falta que los nombre, ellos saben quienes son y lo importantes que son para mi y, aunque algunos están lejos, tengo la suerte de poder sentirme siempre acompañado.

A las personas que, aunque no aparecen aquí con nombres y apellidos, han estado presentes de alguna forma durante el desarrollo de este trabajo y han hecho posible que hoy vea la luz.

A todos mi eterno agradecimiento.

Resumen

Durante los últimos años, en el área de la Interacción Humano-Robot (HRI), ha sido creciente el estudio de la interacción en la que participan usuarios no entrenados tecnológicamente con sistemas robóticos. Para esta población de usuarios potenciales, es necesario utilizar técnicas de interacción que no precisen de conocimientos previos específicos. En este sentido, al usuario no se le debe presuponer ningún tipo de habilidad tecnológica: la única habilidad interactiva que se le puede presuponer al usuario es la que le permite interaccionar con otros humanos. Las técnicas desarrolladas y expuestas en este trabajo tienen como finalidad, por un lado que el sistema/robot se exprese de modo y manera que esos usuarios puedan comprenderlo, sin necesidad de hacer un esfuerzo extra con respecto a la interacción con personas. Por otro lado, que el sistema/robot interprete lo que esos usuarios expresen sin que tengan que hacerlo de modo distinto a como lo harían para comunicarse con otra persona. En definitiva, se persigue imitar a los seres humanos en su manera de interactuar.

En la presente se ha desarrollado y probado un sistema de interacción natural, que se ha denominado Robotics Dialog System (RDS). Permite una interacción entre el robot y el usuario usando los diversos canales de comunicación disponibles. El sistema completo consta de diversos módulos, que trabajando de una manera coordinada y complementaria, trata de alcanzar los objetivos de interacción natural deseados. RDS convive dentro de una arquitectura de control robótica y se comunica con el resto de sistemas que la componen, como son los sistemas de: toma de decisiones, secuenciación, comunicación, juegos, percepción sensoriales, expresión, etc.

La aportación de esta tesis al avance del estado del arte, se produce a dos niveles. En un plano superior, se presenta el sistema de interacción humano-robot (RDS) mediante diálogos multimodales. En un plano inferior, en cada capítulo se describen

los componentes desarrollados expresamente para el sistema RDS, realizando contribuciones al estado del arte en cada campo tratado. Previamente a cada aportación realizada, ha sido necesario integrar y/o implementar los avances acaecidos en su estado del arte hasta la fecha. La mayoría de estas contribuciones, se encuentran respaldadas mediante publicación en revistas científicas.

En el primer campo en el que se trabajó, y que ha ido evolucionando durante todo el proceso de investigación, fue en el campo del Procesamiento del Lenguaje Natural. Se ha analizado y experimentado en situaciones reales, los sistemas más importantes de reconocimiento de voz (ASR); posteriormente, algunos de ellos han sido integrados en el sistema RDS, mediante un sistema que trabaja concurrentemente con varios motores de ASR, con el doble objetivo de mejorar la precisión en el reconocimiento de voz y proporcionar varios métodos de entrada de información complementarios. Continuó la investigación, adaptando la interacción a los posibles tipos de micrófonos y entornos acústicos. Se complementó el sistema con la capacidad de reconocer voz en múltiples idiomas y de identificar al usuario por su tono de voz.

El siguiente campo de investigación tratado corresponde con la generación de lenguaje natural. El objetivo ha sido lograr un sistema de síntesis verbal con cierto grado de naturalidad e inteligibilidad, multilenguaje, con varios timbres de voz, y que expresase emociones. Se construyó un sistema modular capaz de integrar varios motores de síntesis de voz. Para dotar al sistema de cierta naturalidad y variabilidad expresiva, se incorporó un mecanismo de plantillas, que permite sintetizar voz con cierto grado de variabilidad léxica.

La gestión del diálogo constituyo el siguiente reto. Se analizaron los paradigmas existentes, y se escogió un gestor basado en huecos de información. El gestor escogido se amplió y modificó para potenciar la capacidad de adaptarse al usuario (mediante perfiles) y tener cierto conocimiento del mundo. Conjuntamente, se desarrollo el módulo de *fusión multimodal*, que se encarga de abstraer la multimodalidad al gestor del diálogo, es decir, de abstraer al gestor del diálogo de los canales por los que se recibe el mensaje comunicativo. Este módulo, surge como el resultado de adaptar la teoría de actos comunicativos en la interacción entre humanos a nuestro sistema de interacción. Su función es la de empaquetar la información sensorial emitida por los módulos sensoriales de RDS (siguiendo un algoritmo de detección de actos comunicativos, desarrollado para este trabajo), y entregarlos al gestor del diálogo en cada turno del diálogo.

Para potenciar la multimodalidad, se añadieron nuevos modos de entrada al sistema. El sistema de localización de usuarios, que en base al análisis de varias entradas de información, entre ellas la sonora, consigue identificar y localizar los usuarios que rodean al robot. La gestión de las emociones del robot y del usuario también forman parte de las modos de entradas del sistema, para ello, la emoción del robot se genera mediante un módulo externo de toma de decisiones, mientras que la emoción del

usuario es percibida mediante el análisis de las características sonoras de su voz y de las expresiones de su rostro. Por último, otras modos de entrada incorporados han sido la lectura de etiquetas de radio frecuencia, y la lectura de texto escrito.

Por otro lado, se desarrollaron nuevos modos expresivos o de salida. Entre ellos destacan la expresión de sonidos no-verbales generados en tiempo real, la capacidad de cantar, y de expresar ciertos gestos "de enganche" que ayudan a mejorar la naturalidad de la interacción: mirar al usuario, afirmaciones y negaciones con la cabeza, etc.

Abstract

In recent years, in the Human-Robot Interaction (HRI) area, there has been more interest in situations where users are not technologically skilled with robotic systems. For these users, it is necessary to use interactive techniques that don't require previous specific knowledge. Any technological skill must not be assumed for them; the only one permitted is to communicate with other human users. The techniques that will be shown in this work have the goal that the robot or system displays information in a way that these users can understand it perfectly. In other words, in the same way they would do with any other human, and the robot or system understands what users are expressing. To sum up, the goal is to emulate how humans are interacting.

In this thesis a natural interaction system has been developed and tested, it has been called Robotics Dialog System (RDS). It allows users and robotic communication using different channels. The system is comprised of many modules that work together co-ordinately to reach the desired natural interactivity levels. It has been designed inside a robotic control architecture and communicates with all the other systems: decision management system, sequencer, communication system, games, sensorial and movement skills, etc. This thesis contributes to the state-of-the-art in two levels. First, in a high level, it is shown a Human-Robot Interaction System (RDS) with multimodal dialogs. Second, in the lower level, in each chapter the specifically designed components for this RDS system will be described. All of them will contribute to the state-of-the-art individually to their scientific subject. Before each contribution it has been necessary to update them, either by integrating or implementing the state-of-the-art techniques. Most of them have been checked with scientific journal papers. The first works were done in the Natural Language Processing system. Analysis and experiments have been carried out with the most important existing voice recognition

systems (ASR) in daily real situations. Then, some of them have been added into the RDS system in a way that they are able to work concurrently, the goal was to enhance the voice recognition precision and enable several complementary input methods. Then, the research focus was move to adapt the interaction between several types of microphones and acoustic environments. Finally, the system was extended to be able to identify several languages and users, using for this later their voice tone.

The next system to be focused was the natural language generator, whose main objectives within this thesis boundaries were to reach a certain level of intelligence and naturalness, to be multilingual, to have several voice tones and to express emotions. The system architecture was designed to be comprised of several modules and abstraction layers because several voice synthesis engines needed to be integrated. A pattern-based mechanism was also added to the system in order to give it some natural variability and to generate non-predefined sentences in a conversation.

Then the Dialog Management System (DMS) was the next challenge. First of all, the existing paradigms whose behaviour is based in filling information gaps were analysed to choose the best one. Secondly, the system was modified and tailored to be adapted to users (by means of user profiling) and finally, some general knowledge was added (by using pre-defined files). At the same time the Multi-modal Module was developed. Its goal is to abstract this multi-modality from the DMS, in other words, the DMS system must use the message regardless the input channel the message used to reach it. This module was created as a result of adapting the communicative act theory in interactions between human beings to our interaction system. Its main function is to gather the information from the RDS sensorial modules (following an ad-hoc communicative act detection algorithm developed for this work) and to send them to the DMS at every step of the communicative process. New modes were integrated on the system to enhance this multi-modality such as the user location system, which allows the robot to know the position around it where the users are located by analysing a set of inputs, including sound. Other modes added to the system are the radio frequency tag reader and the written text reader. In addition, the robot and user emotion management have been added to the available inputs, and then, taken into account. To fulfil this requirement, the robot emotions are generated by an external decision-maker software module while the user emotions are captured by means of acoustic voice analysis and artificial vision techniques applied to the user face. Finally, new multi-modal expressive components, which make the interaction more natural, were developed: the capacity of generating non-textual real-time sounds, singing skills and some other gestures such as staring at the user, nodding, etc.

Índice general

Agradecimientos III

Resumen V

Abstract IX

1. Introducción 1
- 1.1. Planteamiento . . . 2
- 1.2. Aportaciones principales . . . 2
  - 1.2.1. Aportaciones generales del sistema de interacción . . . 2
  - 1.2.2. Aportaciones específicas de cada componente . . . 4
- 1.3. Estructura del documento . . . 5

2. Sistemas de interacción o diálogo 9
- 2.1. Introducción: el diálogo como el proceso de interacción natural . . . 10
- 2.2. Evolución histórica de los sistemas de interacción . . . 11
- 2.3. Los sistemas de interacción en robótica social . . . 14
- 2.4. Resumen . . . 19

3. Robotics Dialog System (RDS): propuesta de un sistema de interacción natural 21
- 3.1. El ecosistema . . . 22
  - 3.1.1. Entorno software . . . 22
  - 3.1.2. Entorno hardware . . . 24
- 3.2. La multimodalidad soportada por el sistema . . . 24

3.2.1. Modos de entrada de información 25
3.2.2. Modos de salida de información 27
3.3. Componentes de RDS 27
3.4. Principales características de RDS 34
3.5. Resumen 35

4. Gestión del diálogo **37**
4.1. Introducción 38
4.2. Gestores de diálogo en robots sociales 39
4.2.1. Estructuras para la gestión del diálogo: arquitecturas y escalabilidad 39
4.2.2. Expresividad y multimodalidad 41
4.2.3. Representación del conocimiento: modelo del mundo y del usuario 42
4.2.4. Conocimiento compartido 44
4.2.5. Comunicación cooperativa y coordinada 46
4.2.6. Contextualización del diálogo 47
4.2.7. Diálogo y aprendizaje 48
4.2.8. Conversación multiparte (varios interlocutores simultáneos) . . 49
4.2.9. Evaluación: escenarios y métricas 50
Cuidado, rehabilitación y otras terapias 50
Guías en museos,centros comerciales, y otros centros públicos . 52
Robots que aprenden de, en, y para el entorno 53
4.3. Contexto del diálogo 53
4.3.1. Aspectos estructurales 53
4.3.2. Aspectos temporales 55
4.4. IDiM: el gestor en RDS de diálogos interpretados 58
4.4.1. IDiM y la arquitectura de control del robot 58
4.4.2. Implementación y gestión del diálogo 60
4.4.3. La fusión multimodal para *IDiM* 62
4.4.4. Atributos semánticos y huecos de información 63
4.4.5. El gestor de diálogo basado en formularios y huecos de información 64
4.5. Implementación de un diálogo 66
4.5.1. Diseño del lenguaje formal 68
4.5.2. Actos comunicativos de expresión 68
4.5.3. Mapa semántico del dominio del diálogo 69
4.5.4. Atributos semánticos, acciones de consulta y acciones de completado 69
4.5.5. Parámetros temporales para el control del diálogo 70
4.6. IDiM respecto a otros gestores del diálogo 70
4.7. Resumen 72

5. Sistema de fusión multimodal **75**
5.1. Introducción: la teoría de actos comunicativos en la interacción entre humanos . . . 76
5.2. Adaptación de la teoría de actos comunicativos al sistema RDS . . . 76
5.3. Cuatro ejemplos de diálogos multimodales . . . 79
5.3.1. El integrador de habilidades por diálogo . . . 79
5.3.2. El Diálogo de teleoperación multimodal . . . 82
5.3.3. Diálogo de reproducción de música en línea . . . 85
5.3.4. Diálogo de pregunta abierta . . . 87
5.4. Resumen . . . 88

6. Sistema de reconocimiento automático del habla **89**
6.1. Introducción . . . 90
6.2. Mecanismos de captura de sonido y voz humana . . . 92
6.3. Dificultades en la interacción por voz: *"Cocktail Party Problem"* . . . 94
6.3.1. Cocktail Party Problem . . . 94
6.3.2. Ruido de ambiente de fuentes diversas . . . 96
6.4. Paradigmas en el reconocimiento automático de voz . . . 98
6.4.1. Basado en gramáticas . . . 99
6.4.2. Basado en un modelo estadístico del idioma (SLM) . . . 100
6.4.3. Basado en un modelo estadístico de un contexto específico . . . 100
6.5. Requisitos para el sistema de reconocimiento de voz efectivo . . . 101
6.6. Análisis de los entornos de reconocimiento de voz disponibles . . . 103
6.7. Integración con la arquitectura de control . . . 106
6.8. Motores de reconocimiento concurrentes . . . 107
6.9. Experimentos sobre la precisión del reconocimiento . . . 111
6.9.1. Utilizando micrófonos auriculares inalámbricos . . . 112
Sin ruido importante de fondo . . . 112
Con ruido de fondo . . . 113
Con ruido cerca del usuario . . . 113
Diferentes volúmenes de voz . . . 114
Diferentes entonaciones . . . 115
Diferentes géneros . . . 115
Diferentes grupos de edad . . . 116
6.9.2. Utilizando micrófono incorporado en el robot (omnidireccional) 116
6.9.3. Utilizando array de micrófonos . . . 118
6.10. Resumen . . . 121

7. Proxémica y sistema de localización multimodal del usuario 123
7.1. Introducción . . . 124
7.2. Localización de la fuente sonora y de usuarios en robótica social . . . 125
7.2.1. El problema de la localización sonora . . . 125
7.2.2. Fonotaxis *vs* proxémica . . . 127
7.2.3. Análisis proxémico en la interacción entre humanos . . . 128
7.2.4. Análisis proxémico en interacción humano-máquina y humano-robot . . . 130
7.3. Factores en análisis proxémico entre el usuario y el robot social Maggie 132
7.3.1. Experiencia de uso . . . 133
7.3.2. La edad . . . 134
7.3.3. La personalidad . . . 135
7.3.4. El género . . . 135
7.3.5. El aspecto del robot . . . 135
7.3.6. El número de usuarios . . . 135
7.3.7. Conclusión: reglas proxémicas observadas en la interacción usuario-Maggie . . . 136
7.4. Descripción del sistema . . . 137
7.4.1. Sistema *hardware*: sensores usados . . . 137
7.4.2. Sistema *software* . . . 140
7.4.3. Integración de la habilidad de localización de usuarios dentro de RDS . . . 142
7.5. Experimentos de localización de usuarios . . . 143
7.5.1. El módulo de localización de fuente sonora . . . 143
7.5.2. El módulo completo de localización de usuarios multimodal . . . 144
7.6. Resumen . . . 145

8. Sistema de detección y gestión de emociones 149
8.1. Introducción . . . 150
8.2. Trabajo relacionado . . . 151
8.2.1. Las emociones a detectar . . . 151
8.2.2. Canales utilizados para la detección emocional . . . 152
8.2.3. Nivel al que se fusiona la información de cada canal . . . 153
8.2.4. Tipos de entrenamiento del sistema . . . 155
8.3. Características del sistema multimodal de detección de emociones propuesto . . . 155
8.3.1. Representación de las emociones a detectar . . . 155
8.3.2. Canales utilizados para la detección de emociones . . . 158
8.3.3. Nivel al que se fusiona la información de cada canal . . . 158
8.3.4. Tipos de entrenamiento del sistema . . . 159

8.4. El sistema de detección por voz propuesto: GEVA 159
8.4.1. La extracción de características sonoras 160
8.4.2. La clasificación por voz 162
8.5. El sistema de detección visión del rostro propuesto: GEFA 165
8.5.1. Detección de la cara 166
8.5.2. Extracción de las características del rostro 166
8.5.3. Clasificación de la expresión del rostro 167
8.5.4. Trabajos completos que agrupan las tres fases 168
8.6. Integración de GEVA y GEFA: sistema completo propuesto 169
8.6.1. Regla de decisión 171
8.7. Experimentos 173
8.7.1. Configuración del experimento 173
8.7.2. Experimentos con GEVA: audio 176
8.7.3. Experimentos con GEFA: visión 177
8.7.4. Experimentos con el sistema completo 178
8.7.5. Rendimiento del sistema completo calculado estadísticamente 180
8.7.6. Rendimiento del sistema de detección de emociones multimodal calculado mediante experimentos con usuarios 181
8.8. Resumen 182

9. Sistema de síntesis de voz con emociones 185
9.1. Introducción 186
9.2. Requisitos 187
9.3. Principales sistemas de síntesis de voz 188
9.4. Descripción del sistema software 190
9.5. Generación de lenguaje natural 193
9.6. Resumen 193

10. Otras modalidades de expresión 195
10.1. Introducción 196
10.2. Los lenguajes musicales 197
10.3. Implementación de la expresión de sonidos no verbales y canto 201
10.4. Gestos expresivos 202
10.5. Resumen 204

11. Conclusiones y trabajos futuros 207
11.1. Conclusiones 208
11.2. Trabajos futuros o en desarrollo 210
11.3. Publicaciones surgidas de este trabajo 212
11.3.1. En revistas científicas 212
11.3.2. En congresos de robótica 213

11.4. Comentarios finales 214

12. Conclusions **215**
12.1. Conclusions 216
12.2. Current and future work 218
12.3. Last comments 220

A. Primeros resultados experimentales con el sistema RDS **221**
A.1. Introducción 222
A.2. Descripción de los experimentos 222
A.2.1. Caso I: interacción individual con adolescentes inexpertos . . . 223
A.2.2. Caso II: interacción en pequeños grupos de niños inexpertos . 224
A.2.3. Caso III: Interacción en pequeños grupos con niños supervisados 225
A.2.4. Caso IV: interacción en grandes grupos con niños supervisados 226
A.3. Análisis de los vídeos 226
A.3.1. Caso I: interacción individual con jóvenes inexpertos 226
A.3.2. Caso II: interacción en pequeños grupos con niños inexpertos . 228
A.3.3. Caso III: interacción supervisada en pequeños grupos de niños 230
A.3.4. Caso IV: interacción supervisada en grandes grupos de niños . 231
A.4. Análisis de los cuestionarios 231
A.4.1. Caso I: interacción individual con adolescentes inexpertos . . . 231
A.4.2. Caso II: interacción en pequeños grupos con niños inexpertos . 233
A.4.3. Caso III: interacción supervisada en pequeños grupos de niños 234
A.4.4. Caso IV: interacción supervisada en grandes grupos de niños . 234
A.5. Conclusiones sobre los experimentos 235

B. Gramáticas y diálogos de ejemplo **237**
B.1. Ejemplos de gramáticas usadas 238
B.1.1. Gramática de bienvenida 238
B.1.2. Gramática Integrador 238
B.1.3. Gramática para la teleoperación multimodal del robot 240
B.1.4. Gramática para la selección de juegos 240
B.1.5. Gramática para el control de la televisión 241
B.2. Diálogos de voz siguiendo el estándar VoiceXML 241
B.2.1. Diálogo de entrada al sistema 241
B.2.2. Diálogo de registro en el sistema 246

C. Glosario de Acrónimos **249**

Bibliografía **253**

Índice de Tablas

6.1. Comparación de los principales entornos de reconocimiento de voz (2011) 105
6.2. Tabla de definiciones 110

8.1. Matrices de confusión para GEVA 176
8.2. Matrices de confusión para GEFA 178
8.3. Matriz de confusión del sistema completo calculada estadísticamente . 180
8.4. Matriz de confusión del sistema completo calculada mediante experimentos 182

9.1. Resumen comparativo de diferentes motores de síntesis de voz 189

10.1. Comparación de los principales lenguajes de programación musical . . 199
10.2. Detalles técnicos de los lenguajes de programación musical 201

A.1. Resultados tras en análisis de los vídeos del Caso I 227
A.2. Valores relativos interesantes del análisis de los vídeos para el Caso I) 227

Índice de figuras

2.1. El sistema de interacción como interfaz 10
2.2. Sistema de diálogo clásico 12
2.3. Sistema de diálogo multisonido 15
2.4. El robot Biron 17
2.5. El robot HRP2 18

3.1. Contexto del Sistema de Diálogo Robótico (RDS) 23
3.2. Robots sociales del RoboticsLab 25
3.3. RDS: Robotic Dialog System 28
3.4. Niveles de Fusión Multimodal 30

4.1. Aspectos estructurales de un proceso de interacción 55
4.2. Aspecto temporal de un proceso de interacción. 57
4.3. El robot social Maggie interactuando 59
4.4. *IDiM* en un vistazo 60
4.5. Cambios de contexto del gestor de diálogo 61
4.6. La fusión multimodal y el gestor de diálogo 62
4.7. Gramática para controlar la televisión mediante voz 64
4.8. Implementación del diálogo de control de la televisión 67

5.1. Actos comunicativos del diálogo 77
5.2. Etiquetas de radio frecuencia con pictogramas 78
5.3. Flujo del diálogo 80
5.4. Capturas del diálogo de teleoperación 84
5.5. Esquema del diálogo de teleoperación 84
5.6. Ejemplo de la fusión multimodal de dos actos comunicativos 86

6.1. Tipos posibles de micrófonos 92
6.2. Fuentes de audio 95
6.3. Cocktail Party Problem 95
6.4. Cancelación activa del eco 97
6.5. Mezcla a la entrada y separación a la salida de los canales de audio . 98
6.6. Arquitectura del sistema de reconocimiento de voz 106
6.7. Tasa de acierto vs SNR: función de probabilidad de tasa de acierto (PSR) 110
6.8. Confianza media vs típica confianza 110
6.9. Receptor-emisor: micrófonos inalámbricos Sennheiser 111
6.10. Resumen de la precisión de reconocimiento usando micrófonos auriculares 114
6.11. Precisión del reconocimiento con distintos volúmenes de voz 115
6.12. Micrófono omnidireccional MP33865 117
6.13. Micrófono de array 118
6.14. Array de micrófonos en la base del robot 119
6.15. Precisión de reconocimiento usando distintos tipo de micrófonos . . . 120

7.1. Espacios personales en la interacción humano-humano según Lambert 130
7.2. Diferencia de aspecto externo entre un robot humanoide y uno mecatrónico 131
7.3. Interacciones con niños de entre 8 y 10 años 134
7.4. Interacciones con personas mayores de 10 años 134
7.5. Interacción grupal 136
7.6. Niños imitando al robot bailando 136
7.7. Reglas proxémicas 138
7.8. Micrófonos en el robot Maggie 139
7.9. Esquema de situación de los micrófonos en Maggie 139
7.10. Localización de usuarios en el sistema de diálogo multimodal 143
7.11. Áreas de localización 145
7.12. Fases de la localización respecto al usuario 146

8.1. Nivel en el que se realiza la fusión 154
8.2. Sistema de detección de emociones multimodal implementado 156
8.3. Esquema del proceso de toma de decisión para determinar la emoción del usuario en cada acto comunicativo 171
8.4. Los robots empleados en los experimentos 174
8.5. Imagen tomada durante los experimentos llevados a cabo en el IST de Lisboa 175

9.1. Arquitectura del sistema de síntesis de voz 190

10.1. Interfaz para construir canciones sintéticas mediante Vocaloid 203

10.2. Robot social Maggie realizando gestos de enganche para el diálogo . . 203

A.1. Interacción individual con adolescentes sin experiencia previa con el sistema 224
A.2. Grupo de niños interactuando con el robot sin ayuda 225
A.3. Investigador presentando a Maggie 225
A.4. Interacción en grandes grupos con niños supervisados 226
A.5. Tarjetas con pictogramas y etiquetas de radio frecuencia 229
A.6. Niños jugando al juego de los peluches 231
A.7. Caso I 232
A.8. Caso II 233
A.9. Caso III 235
A.10. Caso IV 236

CAPÍTULO 1

Introducción

"Las mayores innovaciones del siglo XXI vendrán de la intersección de la biología y la tecnología. Una nueva era está comenzando." — Steve Jobs

1.1. Planteamiento

La interacción humano-robot constituye un campo de investigación amplio. Es en este campo de investigación donde se enmarca el contenido de esta tesis doctoral. Este trabajo de investigación surge como respuesta al problema de que no existe un sistema suficientemente avanzado como para plantear la interacción humano-robot de una manera similar a como se plantea la interacción humana. Por ello, *se persigue el objetivo de crear un sistema que posibilite la interacción entre los robots y los humanos de la manera más natural posible*. Para lograrlo, ha sido necesario llevar a cabo tareas de investigación, desarrollo, e integración en diversos campos relacionados con la interacción humano-robot: diálogos, multimodalidad, expresividad, computación afectiva, etc.

Debido a la complejidad de la interacción natural entre humanos y robots sociales, que involucra muchos campos de investigación, *resulta difícil encontrar un sistema general que cubra todos los aspectos que conciernen a la interacción humano-robot*. Es a este desafío donde esta tesis pretende dar una respuesta, creando un sistema que facilita la interacción entre ambas partes sin la necesidad de ayuda externa.

Se considera, por parte del autor, que con el trabajo aquí expuesto se da un paso en mejorar la interacción humano-robot en pos de lograr una mayor naturalidad, "humanizando la interacción". Se ha puesto especial hincapié, en que el sistema expuesto funcione de manera robusta y sin la necesidad de personal de apoyo. En este sentido, este trabajo logra que una vez que el robot se enciende funcione correctamente el sistema, sea cual sea el usuario y el tiempo de interacción empleado. En un futuro cercano, se espera que el sistema genérico, aquí presentado, constituya la base para la interacción en robots asistenciales.

1.2. Aportaciones principales

A continuación se describen las principales aportaciones realizadas con el desarrollo de este trabajo. Se comienza con las aportaciones realizadas por el sistema general de interacción RDS. Posteriormente, se enumeran las aportaciones realizadas en cada uno de los componentes concretos del sistema, en los que se considera que se ha realizado una aportación relevante.

1.2.1. Aportaciones generales del sistema de interacción

- Sistema de interacción general, llamado RDS, capaz de trabajar con múltiples modos de interacción tanto a la entrada del sistema (fusión de la entrada multimodal), como a la salida expresiva del mismo (fisión multimodal). Entre estos

modos, destacan principalmente los relacionados con el audio (*sistema multisonido*), siendo capaz de aprovechar la entrada de información multimodal para las siguientes tareas: reconocimiento de voz, identificación del usuario, localización espacial del usuario respecto al robot, detección de emociones del usuario; mientras que los canales de salida se información se usan para: síntesis de voz con emociones, síntesis de sonidos no verbales y generación musical. Al modo sonoro (*multisonido*) se lo complementa con el resto de modos para completar la multimodalidad: sistema visual, gestual, de radio frecuencia, táctil. Los primeros resultados han sido publicados en [Alonso-Martin et al., 2013a].

- La entrada multimodal del sistema se gestiona inspirándose en la *teoría de actos comunicativos* que tiene lugar en la interacción entre humanos. De esta manera, se abstrae al "gestor de la interacción" de los modos/canales mediante los que se obtiene el mensaje comunicativo. Cada *acto comunicativo*, consiste en un paquete de información formado por atributos-valores relativos al diálogo multimodal. Se ha desarrollado el componente capaz de agrupar estos valores semánticos de manera coherente temporalmente en paquetes de información (fusión multimodal) correspondientes a cada turno del proceso comunicativo. Publicado en [Alonso-Martín et al., 2013].

- El sistema de interacción es adaptable al usuario. Esta adaptación se realiza mediante el uso de perfiles de usuario que el sistema es capaz de generar y actualizar mediante el propio sistema de interacción natural. La adaptación consiste en:

  1. Adaptación al idioma (multilingüismo): el sistema es capaz de hablar en el idioma con el que se expresa el usuario. Esta adaptación es válida para cualquier idioma reconocido, pero trabaja especialmente bien para español e inglés. Basta con que el usuario salude al sistema para detectar la identidad del usuario y el idioma usado en la comunicación. Publicado en [Alonso-Martin et al., 2011].
  2. Adaptación de la distancia de interacción (proxémica): el sistema es capaz de determinar cual es la distancia de interacción a mantener con cada usuario en concreto. Publicado en [Alonso-Martín et al., 2012].
  3. Adaptación a la familiaridad con el sistema: dada la experiencia de uso del usuario con el sistema (en este caso el robot), la interacción puede ser adaptada. De este modo, un usuario novel con el sistema puede ser "tutorizado" por el propio robot en su uso, mientras que un usuario "experimentado" apenas necesita de diálogos de aclaración o de sugerencias.
  4. Adaptación emocional: el sistema puede adaptar la interacción a la emoción detectada en el usuario, así como al propio estado emocional del robot.

- El sistema está integrado en una arquitectura ampliamente extendida en la comunidad científica que trabaja en robótica, ROS[1]. Esto facilita su uso en casi cualquier plataforma robótica que use ROS sin la necesidad de realizar grandes esfuerzos de integración y configuración del sistema aquí propuesto.

1.2.2. Aportaciones específicas de cada componente

Vista las aportaciones generales, se describen las principales aportaciones particulares realizadas en cada componente que forman parte del sistema de diálogo, que a su vez se enmarcan dentro de un campo de investigación independiente[2].

1. Se ha desarrollado un sistema de reconocimiento de voz capaz de trabajar simultáneamente en varios modos: sujeto a gramáticas o de texto libre (ver capítulo 6 y apéndice C). Para ello, se realiza el reconocimiento de la voz usando dos reconocedores en paralelo, es decir en ejecución concurrente. Adicionalmente, se han analizado las configuraciones software-hardware necesarias para realizar esta tarea de manera satisfactoria con robots sociales y los problemas que son necesarios de solventar. Publicado en [Alonso-Martin & Salichs, 2011].

2. Se ha desarrollado un sistema multimodal de localización de usuarios respecto al robot y se ha integrado dentro del sistema de diálogo. Con esto y un conjunto de reglas (obtenidas tras un estudio proxémico de los usuarios con el robot[3]), se logra adaptar la distancia de interacción a cada usuario en concreto. Publicado en [Alonso-Martín et al., 2012].

3. Partiendo de una tesis anterior y del estándar VoiceXML, se ha extendido la interpretación que se hace del mismo, para soportar todas las mejoras introducidas. De esta manera, se permite en la propia especificación del diálogo el uso del multimodo, del multilingüismo, consultas a web semánticas para responder a preguntas generales, uso de servicios de reproducción de música en-linea, etc.

4. Se ha desarrollado un sistema de síntesis verbal y no verbal, capaz de sintetizar voces y sonidos "con emociones". El sistema verbal desarrollado funciona de manera distribuida, con múltiples motores de síntesis de voz al mismo tiempo (actualmente cuatro) y gestionar temporalmente la cola de locuciones, con funciones de síntesis tales como: pausar/reanudar la locución en curso, cancelar, cambiar la emoción, "secuestrar" la síntesis sonora, generar sonidos no verbales de síntesis en tiempo real que "concuerdan" con los sonidos percibidos, etc.

[1] `http://www.ros.org/`

[2] Muchas de estas aportaciones se encuentran refrendadas por su publicación en revistas y conferencias

[3] En este caso concreto estás reglas se han obtenido con el robot Maggie

El sistema de síntesis de sonidos no verbales es capaz de expresar sonidos que simulan emociones del propio robot o replicar la voz del usuario en el propio lenguaje sonoro del robot[4]. Publicado en [Alonso-Martin et al., 2012].

5. Se ha desarrollado un sistema de detección de emociones multimodal integrado con el sistema de interacción. El sistema usa como entrada de información: la detección emocional por análisis de la voz del usuario y el análisis de las expresiones de su rostro mediante técnicas de visión artificial. Esta detección emocional se puede usar para adaptar los diálogos en su especificación en ficheros VoiceXML. Publicado en [Alonso-Martin et al., 2013b].

1.3. Estructura del documento

Esta tesis comienza con una introducción sobre el proceso de interacción natural en la comunicación entre humanos. Se presenta una descripción de los estudios realizados en ese campo, en los que se introduce al diálogo como un proceso de interacción multimodal. Seguidamente se revisa el estado del arte sobre los trabajos que tienen el objetivo de adaptar la manera de interactuar entre los humanos, para llevarlo al ámbito de la comunicación humano-máquina, con especial interés en el proceso comunicativo en el que la máquina es un robot con capacidades de interacción sociales. A continuación, se presenta el sistema de interacción desarrollado en este trabajo, denominado RDS (Robotic Dialog System). Este sistema se enmarca dentro de una arquitectura de control de robots sociales y se detallan sus principales características. Una vez presentado RDS, en cada capítulo se describen los principales componentes desarrollados expresamente por el doctorando para este trabajo, y que forman parte del sistema RDS. Se finaliza con los resultados experimentales llevado a cabo con usuarios que sirven para dar validez a este sistema.

Es importante remarcar que **se ha seguido una aproximación "top-down"** (de arriba hacia abajo), de tal manera que primero se presenta el sistema de interacción general desarrollado, y posteriormente se describirse en detalle cómo se ha resuelto cada uno de los sub-problemas que conforman el problema general de interacción natural y que por sí mismos, se pueden consideran un campo de investigación propio. Se ha seguido esta aproximación para primero mostrar una visión general del sistema que permita comprender la integración de cada módulo con el sistema general. Por supuesto, se podía haber optado por un orden diferente a la hora presentar los capítulos (cronológico, de abajo hacia arriba, siguiendo el flujo de la información, etc), pero se ha considerado que este es el más adecuado.

En los sucesivos capítulos se describen los componentes que forman parte del sistema. De esta manera, se han adaptado las técnicas más recientes al sistema RDS,

[4]Recordar al robot R2D2 de *La guerra de las galaxias*

además de realizar alguna aportación significativa al avance del estado del arte en cada campo de investigación abordado. Estas aportaciones, en su mayoría se han visto reflejadas en forma de publicación científica. Siguiendo esta aproximación, *cada capítulo tienen su propio estado del arte, aportaciones y resumen.*

El contenido es explicado en capítulos del siguiente modo:

- **Capítulo 1**. Corresponde con el capítulo actual y en él se introduce al lector en el problema tratado por esta tesis, así como las soluciones adoptadas. Se enumeran las principales aportaciones realizadas y la estructura interna de este documento.

- **Capítulo 2**. En este capítulo se establecen los principales conceptos que inspiran el resto del documento. Se introduce al lector el problema de la interacción natural entre humanos, y entre humanos y máquinas. Se describen los estudios realizados, que investigan la interacción natural entre los humanos, para posteriormente realizar un análisis exhaustivo del estado del arte en la interacción humano-máquina y especialmente en la interacción humano-robot.

- **Capitulo 3**. En este capítulo se presenta el sistema de interacción propuesto: RDS (Robotic Dialog System). Este sistema queda enmarcado dentro de la arquitectura software de control AD-ROS, usada para el control de todos los sistemas del robot, no sólo los de interacción. Se muestra un diagrama general de los componentes que conforman el sistema RDS y de su relación entre ellos. Seguidamente se detallan las principales funciones de cada componente y las características generales que identifican al sistema RDS.

- **Capitulo 4**. En este capítulo se comienza con la descripción del primero de los subsistemas que conforman el sistema general de interacción, es el encargado de la gestión del diálogo. Se realiza un estudio del estado del arte pormenorizado, con los principales gestores de diálogos agrupados en categorías. Seguidamente se presenta nuestro gestor de diálogo, IDiM (*Interpreted Dialog Management*), basado en el relleno de huecos de información. Se describe como implementar diálogos usando este gestor y se compara con el resto de gestores analizados.

- **Capítulo 5**. En este capítulo se muestra cómo se realiza el proceso de fusión multimodal, que logra abstraer al gestor del diálogo de los canales (modos) usados para la extracción de la información que subyace en cada turno del diálogo. Esta manera de fusionar la información adapta la *teoría de actos comunicativos de la interacción humana* al entorno de la interacción humano-robot. Para su compresión, se describen y detallan los principales diálogos implementados en esta tesis.

- **Capítulo 6**. Aquí se muestra cómo se ha integrado, dentro del sistema de interacción, el proceso de reconocimiento automático del habla. Se ha realizado un estudio exhaustivo de las diversas configuraciones software/hardware que posibilitan esta tarea y de los desafíos que actualmente están por resolver. Se presenta la solución adoptada y experimentos sobre la misma.

- **Capítulo 7**. En este capítulo se trata el problema de la localización de los usuarios espacialmente respecto al robot. Se analiza cómo se ha resuelto en la literatura este problema. Seguidamente mediante experimentos con usuarios (estudios proxémicos), se comprueba como se sitúan respecto al robot social Maggie, en función de diversos aspectos, para establecer reglas válidas para nuestro sistema de interacción. Con estas reglas y el sistema bimodal de localización de usuarios (sonora y visual), se logra que el sistema pueda localizar espacialmente al robot en una situación adecuada al contexto comunicativo.

- **Capítulo 8**. En este capítulo se presenta la detección y gestión de las emociones y el género dentro del sistema de interacción. Se realiza un estado del arte de las metodologías y técnicas usadas. Se presenta nuestro subsistema de detección de género y emociones multimodal basado en: voz y rasgos faciales. Se ha desarrollado un módulo de detección de género y emociones por voz, además de integrado dos sistemas comerciales (CERT y SHORE) de detección de emociones y género mediante visión del rostro del usuario. Sobre estas tres fuentes de información se toma una decisión de la emoción predominante en el usuario y su género, esta información se incorpora como entrada multimodal en el acto comunicativo correspondiente.

- **Capítulo 9**. Con este capítulo se comienza con la descripción de los componentes expresivos del sistema, en concreto con el de la síntesis de voz. Se hace un análisis de las soluciones posibles. Se presenta el sistema de síntesis de voz, capaz de sintetizar voz con emociones y con varios motores de síntesis. Finalmente se describe sus características y como ha sido introducido dentro del sistema de interacción.

- **Capítulo 10**. En este capítulo se describe el resto de componentes expresivos del sistema como son: los gestos, los sonidos no verbales, y la capacidad de cantar. Cada uno de ellos se describe individualmente y su integración con el sistema. Se proponen nuevas vías de desarrollo en cada uno de ellos.

- **Capítulo 11**. En este capítulo final se reflexiona sobre el trabajo realizado a modo de conclusiones y se establecen las futuras líneas de trabajo en la mejora del sistema de interacción natural planteado.

- **Apéndice 1**. Experimentos realizados con usuarios reales, en una fase intermedia en el ciclo de desarrollo de esta tesis. Provocó la introducción de importantes mejoras en la interacción.
- **Apéndice 2**. Código fuente de algunos componentes específicos del sistema de diálogo como son gramáticas, diálogos VoiceXML, etc. que facilitan la comprensión del sistema y del texto.
- **Apéndice 3**. Constituye el glosario de acrónimos usados durante la redacción del manuscrito.

  Dado que la presente tesis se enmarca dentro de tesis con mención internacional, el texto que aparece en las figuras está en inglés, para facilitar la comprensión del mismo por los revisores expertos de habla no castellana.

CAPÍTULO 2

Sistemas de interacción o diálogo

"El dialogo es el mejor arte en una sociedad, sin él no existiría."— anónimo

2.1. Introducción: el diálogo como el proceso de interacción natural

La RAE define el diálogo como "La plática entre dos o más personas que *alternativamente* muestran sus *ideas* o afectos". Otra definición que se puede encontrar es la que da la Wikipedia: "El diálogo es una forma oral y escrita en la que se comunican dos o más personajes en un intercambio de ideas *por cualquier medio*". Finalmente, en *WordReference*, se encuentra la siguiente definición: "Conversación entre dos o más personas que *intercambian el turno de palabra*".

Como intuitivamente sabemos desde que somos niños, la acción de conversar es el modo más natural para resolver un gran número de acciones cotidianas entre los humanos. Por otro lado, es una herramienta clave para lo que se ha venido a denominar "interacción natural", o "natural interaction"(NI) en su traducción al inglés. Los estudios en NI se centran en en el análisis de las características esenciales y descriptivas de la interacción entre humanos. Estos estudios no resultan sencillos, primero porque el potencial del lenguaje es mucho mayor que un mero mecanismo de "intercambio de información", como es normalmente se asume. Segundo, porque la cantidad de información intercambiada es a menudo mayor que la información adquirida (por el receptor).

El objetivo es disponer de sistemas que faciliten la comunicación persona-máquina del modo más natural posible, es decir, a través de la conversación. Se persigue imitar a los seres humanos en su manera de actuar de tal manera que el sistema de interacción actúe como un intermediario entre el humano y la máquina (ver Fig. 2.1).

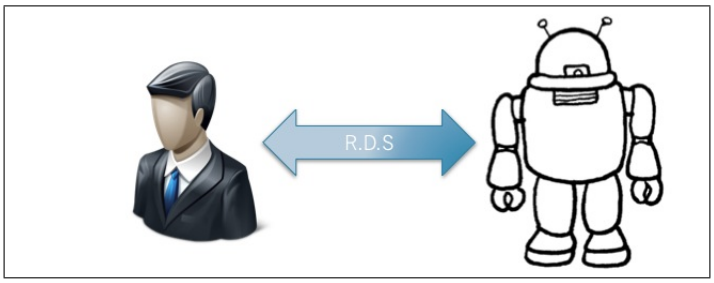


Figura 2.1: El sistema de interacción actúa como una interfaz, o intermediario, entre el humano y la máquina

.

2.2. Evolución histórica de los sistemas de interacción

La creencia de que los humanos seremos capaces de interactuar con las máquinas en una conversación natural ha estado presente como un tema recurrente desde los inicios de la ciencia ficción, recordar por ejemplo: (*"2001: Una odisea del espacio"* (1968), *"Blade Runner"* (1982), o *"Yo, Robot"* (2004). Con las recientes mejoras en el campo de la interacción natural por voz, tales sistemas están empezando a ser factibles. El diálogo es el modo más natural para resolver un gran número de acciones cotidianas entre los seres humanos: obtener una determinada información, contratar un servicio, conocer el estado de un determinado proceso, etc.

Los sistemas de interacción son comúnmente conocidos como sistemas de diálogo, entendiendo el diálogo como el proceso de comunicación entre varias partes, que involucra no sólo al lenguaje como mecanismo de interacción, sino también, gestos, sonidos, emociones, etc. Dichos sistemas y más en concreto los sistemas multimodales, son un área activa de investigación.

La interacción por diálogo no es un tema de investigación nuevo, y existen numerosas aproximaciones al problema del diálogo natural. Varias investigaciones se centran en dicho problema, por citar algunos se pueden leer ([Bach & Harnish, 1979], [Bruner, 1975], [Searle, 1975, Searle, 1969]). Sin embargo, los sistemas de diálogo más desarrollados se han implementado en contextos de comunicación muy restrictivos, como el teléfono o sistemas de acceso mediante PC. La implementación de sistemas de diálogo dentro de la robótica no sigue el mismo nivel de desarrollo. Esto es debido, principalmente, a que los robots tienen cuerpo físico y se mueven por el entorno natural. Los sistemas de diálogo tienen que estar coordinados con los sistemas de percepción y actuación dentro de un entorno real y desestructurado, siendo una tarea mucho más compleja que la realizada para entornos de atención al cliente telefónicos o para controlar un smartphone. Sin embargo, es precisamente en estos entornos de interacción con robots, donde la multimodalidad cobra su mayor importancia ([Cheyer et al., 1998], [Gorostiza et al., 2006a], [Niklfeld et al., 2001], [Seneff et al., 1996], [Wahlster, 2003a], [Waibel & Suhm, 1997]).

Beneficiándose de las interesantes mejoras en las áreas del reconocimiento del habla, el procesamiento del lenguaje natural y de la síntesis del habla, las primeras iniciativas de investigación relacionadas con los sistemas de diálogo oral surgen a principios de los años 80. El origen de esta área de investigación está ligado a dos grandes proyectos: el programa DARPA Spoken Language System en E.E.U.U y Esprit SUNDIAL en Europa. DARPA se centraron en usar las tecnologías del habla al dominio de la reserva de vuelos vía telefónica ([Walker et al., 2001]). El proyecto SUNDIAL trataba con información sobre horarios de avión o tren en cuatro idiomas europeos. Esta investigación fue el origen de otros numerosos proyectos financiados

por la Comunidad Europea relativos principalmente al modelado del diálogo, por ejemplo VERMOBIL, DISC, ARISE, CSELT, LIMSI, etc.

Un sistema de diálogo oral tradicional suele ser un sistema que, para gestionar de forma exitosa la interacción con los usuarios, incorpora cinco módulos: reconocimiento automático del habla (*Automatic Speech Recognition* o *ASR*), entender el lenguaje natural (*Natural Language Understanding* o *NLU*), gestión del dialogo (*Dialog Manager* o *DM*), generación del lenguaje natural (*Natural Language Generation* o *NLG*) y conversión de texto a voz (*Text to Speech* o *TTS*). Ver fig 2.2.

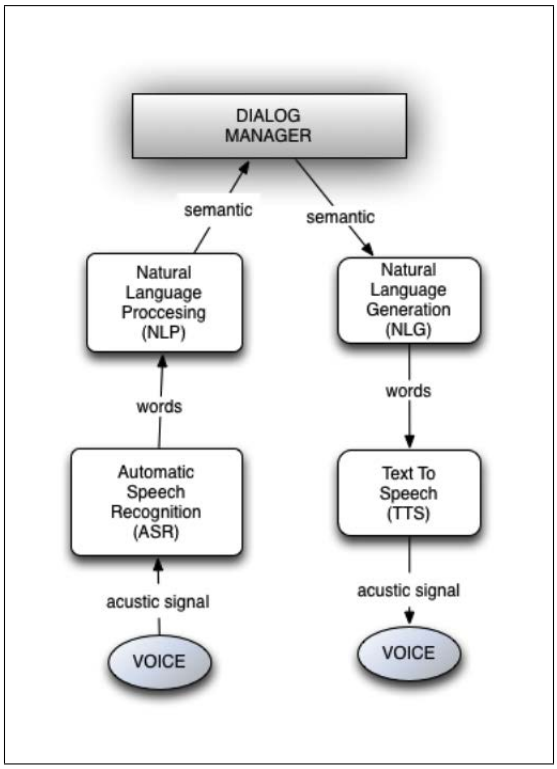


Figura 2.2: Esquema clásico de un sistema de diálogo sin multimodalidad. Este tipo de sistemas se usa frecuentemente en sistemas de atención automática de llamadas (*call centers*). La voz representa el único modo de entrada y salida de información del sistema.

En los años 90 se introduce la capacidad de multimodo, en ese sentido destaca *Communicator*. Este proyecto, que forma parte de la iniciativa DARPA, tuvo como objetivo el desarrollo de tecnologías del habla novedosas que pudiesen emplear como entrada no sólo la voz sino también otras modalidades. Los sistemas desarrollados en E.E.U.U y Europa eran capaces de interactuar con los usuarios en múltiples dominios, en los cuales el usuario y el sistema podían iniciar la conversación, cambiar

de tema o interrumpirse. En el se basaron otros proyectos como el *Carnegie Mellon Communicator* ([Rudnicky & Thayer, 1999]). Las líneas de investigación durante estos años estuvieron relacionadas con la mejora de las tasas de éxito de los diversos módulos de los sistemas de diálogo, sobre todo en la precisión del reconocedor de voz y la independencia del locutor. A finales de los años 90 y con la proliferación de la telefonía móvil, la investigación puso el foco en mejorar la interacción en ambientes muy ruidosos ([Lamel, 2002],[Gauvain, 2002]), y de manera general en aumentar la robustez de los módulos anteriormente presentados.

A partir de 2003, los expertos han propuesto objetivos de más alto nivel, tales como proveer al sistema de razonamiento avanzado (mediante planificación, agentes, sistemas de aprendizaje automático), facultad de adaptación, proactividad, multimodalidad y multilingüismo. Estos nuevos objetivos hacen referencia al sistema en su conjunto y representan tendencias importantes que se alcanzan, en la práctica, a través del trabajo común en diversas áreas y diversos componentes del diálogo. Así en contraste con la década anterior cuando los objetivos se definieron por cada área (módulo), las tendencias actuales de investigación se definen en grandes objetivos compartidos.

La **proactividad** es necesaria para que las máquinas pasen de ser consideradas herramientas y se conviertan en verdaderas entidades conversacionales. Los sistemas proactivos tienen la capacidad de entablar una conversación con el usuario incluso cuando este aún no haya solicitado explícitamente la intervención del sistema.

El interés por el desarrollo de sistemas con los cuales obtener una conversación tan natural y rica como la producida entre seres humanos, fomenta la investigación sobre **interfaces multimodales**. En ellas, al contrario que con el uso de las interfaces tradicionales existe flexibilidad en los modos de entrada y de salida. Los primeros sistemas de diálogo multimodales combinaban el habla con mapas gráficos ([Cheyer et al., 1998]) o con escritura ([Waibel & Suhm, 1997]) . La mayoría de estos sistemas usan la multimodalidad a la entrada o a la salida del sistema. Es en la década de 2000 cuando nuevos proyectos como *Smartkom* ([Wahlster, 2003a, Wahlster, 2001a, Wahlster, 2001b, Wahlster, 2006, Reithinger & Alexandersson, 2003]) se han dirigido en lograr la multimodalidad tanto en la entradas del sistema como en la salidas del sistema, lo que se conoce como **simetría multimodal**. El citado proyecto *Smartkom*, con numerosas publicaciones, patentes y productos comerciales, introduce un sistema de diálogo en el que la simetría multimodal es su principal virtud. En este sistema se habla de fusión multimodal de las entradas al sistema de diálogo, y de fisión multimodal, de las posibles salidas del sistema([Wahlster, 2003b]). La aportación de Smartkom no sólo radica en tratar esta multimodalidad de entrada y salida, sino también en los mecanismos de: gestión del diálogo, desambiguación mutua, sincronización de las entradas multimodales, resolución de elipsis multimodal. Sin embargo, su campo de acción no es el de la robótica,

sino interacción causal en aeropuertos, estaciones de tren, hoteles, restaurantes, etc.

La **adaptabilidad** se refiere a la adaptación del sistema de diálogo a cada usuario particular. Los usuarios novatos y los usuarios experimentados pueden desear que el interfaz se comporte de forma diferente. Otro ejemplo de adaptabilidad es el multilingüismo, un mismo sistema de dialogo puede reconocer y expresarse en varios idiomas, en función del idioma del locutor. Tal y como se describe en ([Jokinen, 2003]) , existen dos niveles en los que el sistema puede adaptarse al usuario. El más simple es a través de perfiles de usuario, con las particularidades de cada usuario, idioma preferido, nivel de experiencia actual con el sistema, edad, nombre, interacción sobre su tipo de voz, etc. Un enfoque mas complejo, trata de adaptarse a los aspectos más dinámicos del usuario, como son sus intenciones y el estado emotivo del mismo. Siguiendo esta aproximación, surge la computación afectiva.

Desde 2004, existen numerosos proyectos basados en la **inteligencia emocional** o computación afectiva para mejorar los sistemas de diálogo. Los sistemas de diálogo con inteligencia emocional tratan el comportamiento emocional del usuario para mejorar la interacción entre ambos, reduciendo el número de interrupciones, mal entendidos y en general frustración con el sistema ([Martinovsky, 2006]).

En cuanto a la **portabilidad**, trata sobre la independencia del dominio, del idioma y de la tecnología. Idealmente, los sistemas deberían poder trabajar en diferentes dominios de aplicación, o al menos ser fácilmente adaptables entre ellos.

En este trabajo se ha añadido un nuevo concepto a los ya presentados en la evolución de los sistemas de diálogo. Este concepto es el de **multisonido** (ver la fig. 2.3). Se trata de usar el canal de audio de entrada y salida del sistema para funciones adicionales al modo verbal (entender y generar voz). Entre estas funciones destacan la capacidad de analizar el mensaje verbal para obtener la emoción con la que se comunica el interlocutor, analizar las diferentes señales de audio que se reciben por cada micrófono para intentar localizar el origen espacial de la fuente sonora, analizar el nivel de excitación del entorno, identificar y verificar el interlocutor, expresar sonidos no verbales o generar melodías de voz (capacidad de cantar).

2.3. Los sistemas de interacción en robótica social

Hasta ahora, se ha descrito qué es el diálogo natural y cómo sirve a los humanos para transmitir ideas, conocimientos, sentimientos, etc. por medio de diversos canales de comunicación, siendo el verbal el que cobra mayor importancia. Posteriormente se ha hablado de los sistemas de interacción natural que tratan de imitar la manera en que se comunican los humanos, para adaptarlo a la comunicación humano-máquina. Además se ha visto como estos sistemas han evolucionado a lo largo de las últimas décadas. Sin embargo, no se ha hablado de sistemas de interacción/diálogo aplicados a nuestro campo de investigación: *la robótica social*.

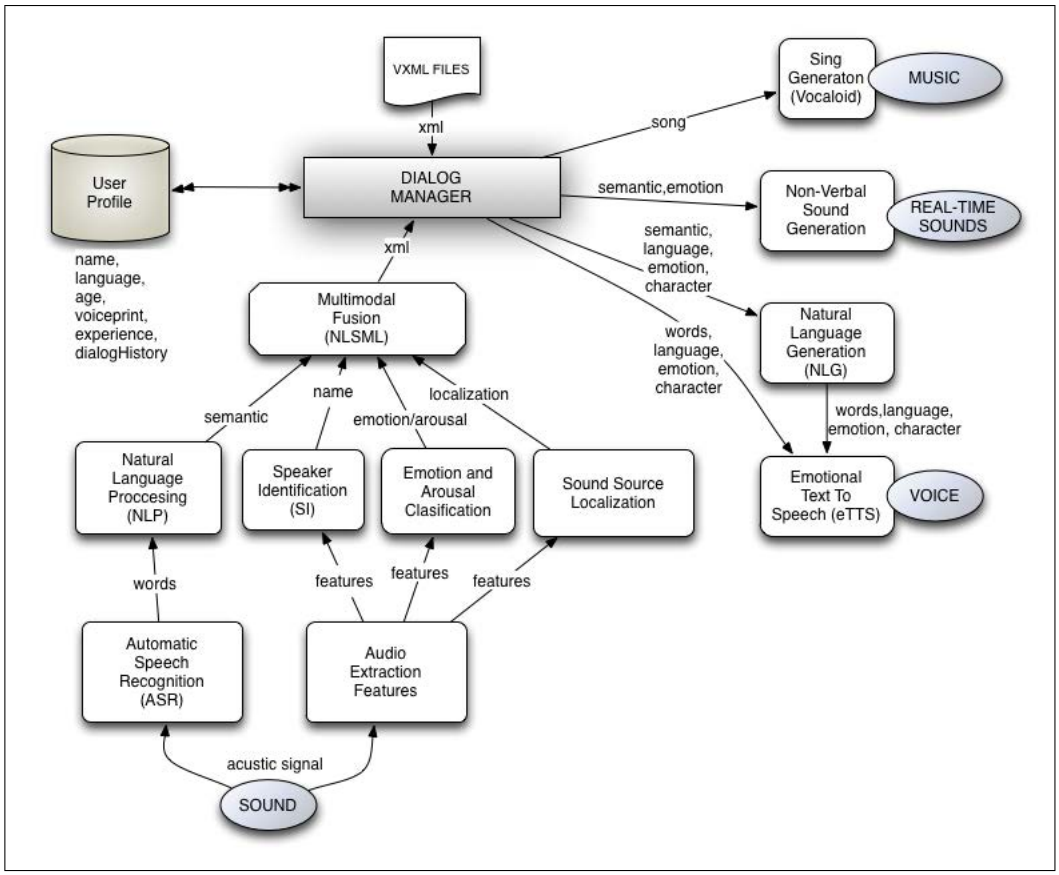


Figura 2.3: Un sistema de diálogo multisonido usa el canal de audio de entrada y salida para más tareas que procesar o generar voz (modo verbal de un sistema de diálogo clásico). Entre estas tareas adicionales, destacan mecanismos para captar emociones, detectar la excitación del entorno, localización de la fuente sonora, detectar la actividad de voz, identificar y verificar el interlocutor, expresión de sonidos no verbales, etc.

En este trabajo, se entiende un un robot social como aquel que interactúa y se comunica con las personas, de forma sencilla y agradable, siguiendo comportamientos, patrones y normas sociales. Por ello, necesita de ciertas capacidades que se enmarcan dentro del dominio conocido como inteligencia social. La mayoría de los robots actuales funcionan relativamente bien en entornos controlados, como – por ejemplo – los laboratorios, pero surgen problemas cuando operan en entornos con condiciones variables, como oficinas, casas o fábricas. No obstante, es de esperar que en poco tiempo el robot social se convertirá en un componente habitual en la vida de cualquier humano, como alguna vez lo hizo el teléfono, la televisión, el automóvil, y la computadora personal.

Como se ha comentado, la implementación de sistemas de diálogo dentro del campo de la robótica no sigue el mismo nivel de desarrollo que los sistemas vistos en la sección anterior. Esto es debido principalmente, a que los robots tienen cuerpo físico y se mueven por el entorno natural. Por tanto, los sistemas de diálogo tienen que estar coordinados con los sistemas de percepción y actuación dentro de un entorno real, variable, y desestructurado, siendo una tarea mucho más compleja que la realizada en laboratorios, en *call-centers* o para el control de un *smartphone*. Como se ha comentado en la introducción, es precisamente en el ámbito de la robótica social, donde la multimodalidad cobra su mayor importancia, debido a la variedad sensitiva y expresiva de los robots sociales, donde el número de sensores y actuadores es más elevado que en cualquier otro sistema [Cheyer et al., 1998] [Gorostiza et al., 2006a] [Niklfeld et al., 2001] [Seneff et al., 1996] [Wahlster, 2003a] [Waibel & Suhm, 1997].

En sentido general, un sistema multimodal soporta la comunicación con el usuario a través de diferentes modalidades como la voz, los gestos, la escritura, sonidos no verbales, etc.[Laurence & Nigay. Coutaz, 1993]. Literalmente, "multi" se refiere a "más de uno" y el termino "modal" cubre la noción de "modalidad" así como la de "modo". Modalidad se refiere al tipo de canal de comunicación utilizado para transmitir o adquirir información. También abarca a la forma en que se expresa o se percibe una idea, o la manera en que se realiza una acción. Nuestra definición de multimodalidad transmite dos características principales que son relevantes para el diseño de software de los sistemas multimodales: la fusión de diferentes tipos de datos desde/hacia diferentes dispositivos de entrada/salida (E/S), y las limitaciones temporales impuestas en el procesamiento de información desde/a dispositivos de E/S.

Hasta la fecha, no se encuentra mucha literatura sobre la implementación de sistemas de dialogo multimodales en HRI. En esta sección, se describen los principales sistemas que se pueden encontrar:

Para el robot *Jijo-2*, en el artículo [Fry et al., 1998] se presenta un sistema de diálogo que únicamente usa la modalidad de la voz como medio de interacción. En otros, más avanzados se introduce el concepto de multimodalidad, por citar algunos: [Cassimatis et al., 2004], [Perzanowski et al., 2001a], y [Lemon et al., 2002]. En

ellos la información es transmitida verbalmente y mediante pantallas táctiles. Ambas fuentes de información complementaria son fusionada de manera que son capaces de resolver frases como: "vete allí", al mismo tiempo que se apunta con el dedo una localización en el mapa de la pantalla táctil. En [Lemon et al., 2002], la interfaz visual es también usada por el sistema de diálogo para mostrar vídeos al mismo tiempo que se sintetiza voz. Por ejemplo, el robot pregunta "¿Es ese el coche rojo que estabas buscando?", mientras se muestra en la pantalla táctil el flujo de imágenes que el robot percibe a través de su cámara.

En [Hüwel et al., 2006] y [Toptsis et al., 2004] se presenta el robot *Biron* (ver Fig. 2.4). Su sistema de diálogo fusiona conjuntamente información verbal con información gestual, siendo capaz de resolver oraciones como: "esto es una planta", mientras se señala con el dedo a la planta. Este tipo de fenómenos, en que el artículo demostrativo (en este caso "esto") es sustituido por el objeto físico que se señala, se le conoce como deíxis verbal[1].

Figura 2.4: El robot Biron es un robot móvil y social capaz de interactuar con las personas. Construido en la universidad de Bielefeld (Alemania) y presentado en 2004 . La interacción es controlada por un sencillo sistema de interacción natural multimodal basado en voz y una pantalla táctil.

Ademas, sistemas con fusión multimodal han sido usados en varios trabajos de

[1]es la parte de la semántica y la pragmática que está relacionada con las palabras que sirven para indicar otros elementos. Palabras como tú, hoy, aquí, esto, son expresiones deícticas, que nos sirven para señalar personas, situaciones, lugares, etc. En pragmática, las expresiones deícticas dependen, para su correcta interpretación, del contexto del hablante, sobre todo del contexto físico, de los elementos extralingüísticos.

"sistemas de programación natural". Por ejemplo, en [Iba et al., 2002] un robot aspiradora es programado usando gestos y voz, por lo que el robot es capaz de fusionar dicha entrada multimodal para inferir acciones que el usuario quiere que realice. Trabajos similares pueden ser encontrados en [Dominey et al., 2007], donde el robot HRP-2 (ver Fig. 2.5) fusiona información visual y verbal para la manipulación de objetos.

Figura 2.5: El robot social HRP-2 integra en su cabeza una cámara y un array de micrófonos que le permiten interactuar con humanos mediante un sistema de interacción bimodal (visión y voz). Fue presentado en 2004 por el instituto técnico AIST de Japón

.

En [Stiefelhagen, 2004], Stiefelhagen presenta un sistema multimodal que usa voz, gestos y el análisis de la orientación de la cabeza del usuario. Los componentes son: reconocimiento de voz, seguimiento de cara y brazos en 3D, reconocimiento de gestos de apuntar, reconocimiento de la pose de la cara, un gestor de diálogo, síntesis de voz, una plataforma móvil y un cámara de visión estéreo. Cada uno de estos componentes es descrito en el artículo conjuntamente con resultados experimentales. El trabajo y los componentes presentados constituyen los bloques básicos para la interacción multimodal. Estos son usados por robot humanoides cooperativos del grupo de trabajo alemán (Sonderforschungbereich).

En los últimos tiempos, otros proyectos se han dirigido hacia el desarrollo pleno de la multimodalidad a la entrada y salida del sistema: (**simetría multimodal**). Como ya se mencionó, por sistema de diálogo con simetría multimodal se entiende, que se gestiona tanto la entrada multimodal como la salida multimodal, no teniendo por qué ser exactamente el mismo número de canales usados en ambos casos. La gestión de la entrada multimodal se conoce como fusión multimodal, mientras que la gestión de la salida multimodal se conoce como fisión multimodal. No obstante, resulta difícil encontrar en la literatura sistemas de interacción con simetría multimodal aplicados al campo de la robótica social.

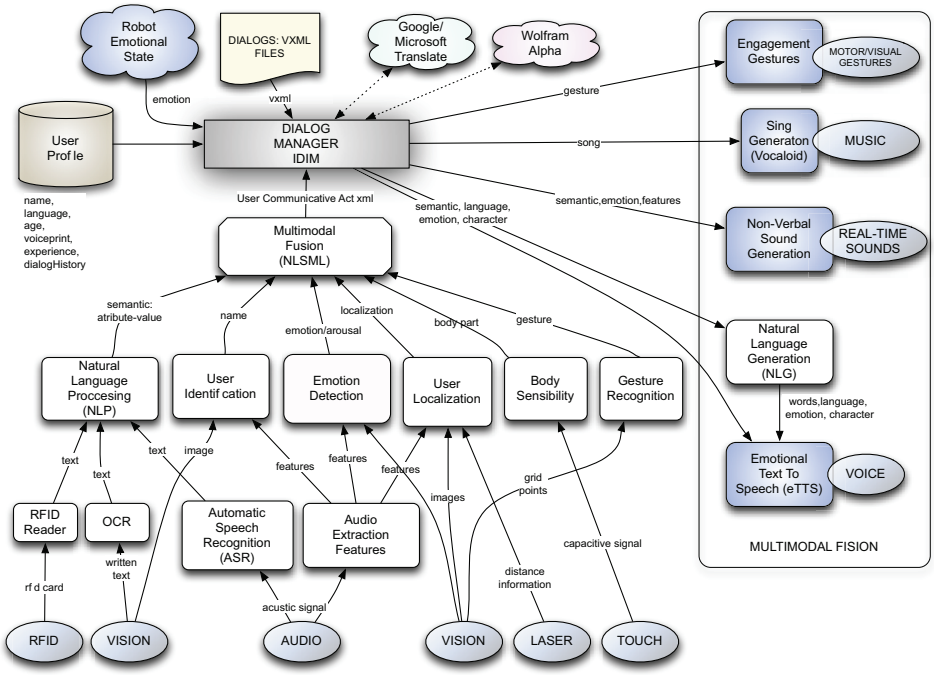
2.4. Resumen

En las dos últimas décadas estos sistemas han ido evolucionando, desde una versión clásica a principios de los años 80, con básicamente cinco módulos (ASR, NLP, DM, NLG, TTS)[2], hasta los modernos sistemas actuales, que incluyen nuevos componentes y multimodo. El uso tradicional de los sistemas de diálogo telefónicos se está desplazando hacia nuevas áreas. Es en su aplicación práctica en el campo de la robótica social, donde el multimodo logra su máxima expresión. Recientes trabajos han supuesto avances en el estado del arte incorporando nuevos componentes, con el objetivo de conseguir interacciones más naturales. Además del multimodo, se han introduciendo mejoras en la interacción relativas a: la proactividad, la adaptabilidad, la gestión de emociones, y la portabilidad.

[2]Ver glosario de términos en el apéndice primero

CAPÍTULO 3

Robotics Dialog System (RDS): propuesta de un sistema de interacción natural



"La mejor forma de predecir el futuro es implementarlo." — David Heinemeier Hansson

3.1. El ecosistema

3.1.1. Entorno software

El sistema de interacción[1] que se presenta a continuación está formado por multitud de módulos que conviven a su vez con otros pertenecientes a la arquitectura de control. La arquitectura de control se encarga de gestionar todas las acciones, tareas, y procesos del robot. Entre estas tareas algunas tienen que ver con la interacción, y por lo tanto con el sistema de interacción, mientras que otras tienen que ver con otros aspectos, como pueda ser la toma de decisión, la exploración del entorno, etc.

Siguiendo un esquema basado en la *modularidad* y la *abstracción* (ver Fig. 3.1) se puede entender el ecosistema en el que funciona el sistema de diálogo. En el nivel inferior se encuentran los componentes hardware: sensores, motores, computadoras, etc. Cada computadora ejecuta un sistema operativo; en el caso de nuestros robots, concretamente la distribución Ubuntu de Linux. A su vez, cada sistema Ubuntu, ejecuta la arquitectura de control robótica. La arquitectura de control es capaz de comunicar los distintos módulos que la conforman, proporcionando los mecanismos de comunicación necesarios. A lo largo de los años que ha durado la elaboración de la tesis doctoral, esta arquitectura de control ha ido evolucionando.

Partiendo de una arquitectura definida teóricamente conocida como AD (arquitectura automática-deliberativa), que fue descrita e implementada en trabajos previos [R. Rivas, A. Corrales, R. Barber, 2007, Barber & Salichs, 2001a], se ha ido evolucionando hasta llegar a la arquitectura actual: una arquitectura híbrida entre AD y el nuevo "estándar de facto" en las arquitecturas robóticas, la arquitectura ROS (Robot Operating System) [Quigley et al., 2009]. Esta nueva arquitectura de control, que se la puede denominar como AD-ROS, permite la convivencia de cualquier módulo desarrollado en "la antigua AD" y "la moderna ROS", de manera que se conserva todo el trabajo desarrollado durante años por nuestro grupo de investigación y se incorpora la posibilidad del uso de los miles de módulos desarrollados por la comunidad ROS. Esto, posibilita que cualquiera de los componentes desarrollados para este trabajo, y del sistema de diálogo completo, puedan ser usados por cualquier investigador, de cualquier lugar del mundo, sobre cualquier robot capaz de ejecutar ROS. La arquitectura de control facilita la tarea de distribuir el software, ya que cualquier módulo puede estar corriendo en cualquier computador, incluso fuera del robot; en este sentido, la única restricción es dada por el módulo (o los módulos) que necesite interactuar directamente sobre un actuador y/o sensor, ya que necesitará comunicarse con él (de manera local o a través de la red).

[1] Recordar que durante todo el documento se usa indistintamente los términos de "sistema de interacción" y "sistema de diálogo", ya que el diálogo es entendido como un proceso de interacción multimodal

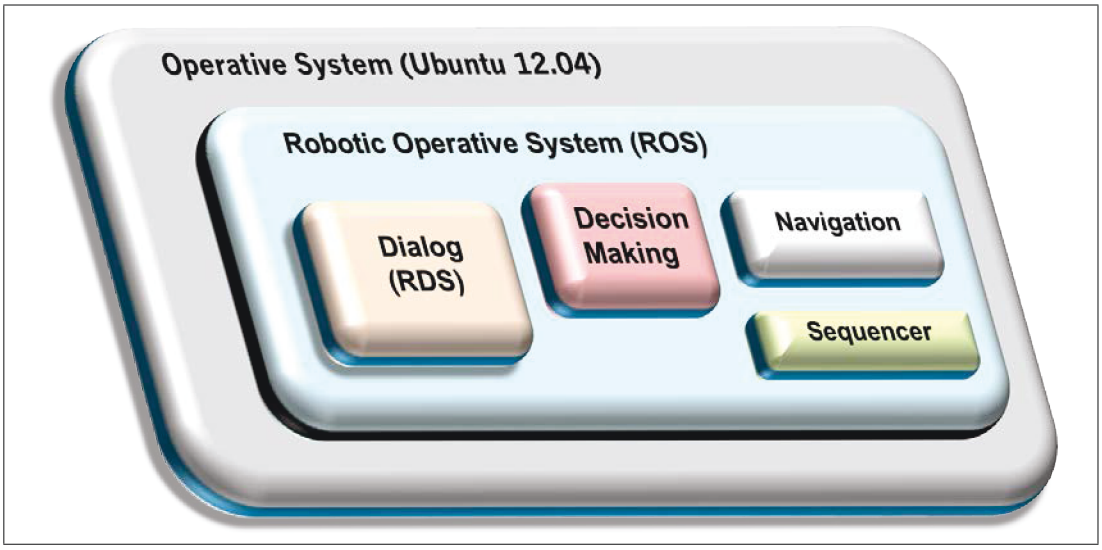


Figura 3.1: El sistema de interacción propuesto (RDS) convive con otros sistemas pertenecientes a la arquitectura de control. En la figura se representa el sistema de interacción RDS, el sistema de toma de decisión ("*Decision Making*"), el sistema de navegación por el entorno ("Navigation"), y el secuenciador ("*Sequencer*"). Estos sistemas se comunican entre sí mediante diversos mecanismos de comunicación que proporciona la arquitectura de control, en este caso la arquitectura ROS. A su vez la arquitectura de control ROS corre sobre sistemas operativos Ubuntu (una de las distribuciones más populares de Linux).

Dentro de la arquitectura robótica AD-ROS conviven multitud de sistemas: el sistema de toma de decisiones, el sistema de navegación, el secuenciador, etc. En ese sentido, el sistema de diálogo es un sistema más dentro de la arquitectura de control, capaz de interactuar con el resto de sistemas, proporcionándoles la capacidad y la abstracción de realizar una interacción de alto nivel con el usuario.

3.1.2. Entorno hardware

Actualmente, el sistema RDS constituye el sistema de interacción de los robots desarrollados por el grupo de trabajo de robots sociales del RoboticsLab[2]. Está siendo utilizado con éxito en los robots *Maggie*, *Mopi*, y *Flory*. Recientemente, se ha incorporado también en dos nuevas plataformas: el robot *Mini*, enmarcado en un proyecto de investigación incipiente en colaboración con la fundación Alzheimer[3] (ver Fig. 3.2), y los robots del proyecto europeo *Monarch*[4].

El robot social Maggie constituye una plataforma de investigación sobre la que han versado numerosos trabajos del grupo, por citar algunos [Gonzalez-Pacheco et al., 2011, Alonso-Martin et al., 2011, Salichs et al., 2006a, Gorostiza et al., 2006b]. En estos trabajos, el robot Maggie ha sido ampliamente descrito. En este trabajo ha sido la plataforma preferida del autor, puesto que consta con el mayor tipo de modalidades de comunicación posibles, además de ser el primero de los robots creados y con el que comenzó el desarrollo de esta tesis.

Como se ha comentado, el sistema también ha sido probado satisfactoriamente en los demás robots sociales del grupo[5]. El robot modular Mopi, es el encargado de dotar de movilidad al robot modular Flory.

Recientemente se está trabajando en la integración del sistema en la nueva plataforma robótica destinada al cuidado y estimulación de personas mayores con problemas de Alzheimer.

3.2. La multimodalidad soportada por el sistema

El sistema RDS está concebido para establecerse como el sistema básico de interacción en cualquier robot que use la arquitectura de control ROS. A continuación se va a detallar cuales son estos canales de entrada y salida que soporta nuestro sistema[6].

[2] http://roboticslab.uc3m.es/roboticslab/

[3] `http://cuidadoalzheimer.com/tag/uc3m/`

[4] `https://mis.u-picardie.fr/R-Discover/images/stories/workshop/sequiera_erf.pdf`

[5] El sistema ha sido probado, pero no al mismo nivel que el robot Maggie. Esto se debe a que se han realizado numerosos shows en conferencias, reuniones, televisiones con el robot Maggie, y no tantos con el resto de robots, que se encuentran actualmente en un nivel de menor madurez

[6] Pese a que el sistema es válido para cualquiera de los robots que presenten algunos o todos estos canales de entrada y salida, esta descripción se centra en el robot Maggie. Sobre esta plataforma

Figura 3.2: Los robots sociales construidos y programados por el grupo de robots sociales de la Universidad Carlos III sirven como plataforma para numerosos trabajos de investigación. Estos robots han servido como soporte para la realización del trabajo aquí presentado, especialmente el robot social Maggie (a la izquierda en la figura)

3.2.1. Modos de entrada de información

- Audio. Para recibir audio, es necesario del uso de uno o varios micrófonos situados estratégicamente. En el robot Maggie, el audio es recibido a través de los micrófonos incorporados en su cuerpo. En la base del mismo, formando una circunferencia de 40 cm de radio y a 21 cm del suelo, se encuentran 8 micrófonos que se conectan por USB al ordenador interno del robot. Estos 8 micrófonos se usan fundamentalmente para tareas de localización de la fuente sonora. Para el análisis de voz, el robot incorpora dos micrófonos en la cabeza. Para una interacción en entornos especialmente hostiles (mucho ruido ambiental), también es posible interactuar con el robot mediante auriculares inalámbricos o bluetooth [7], con la consiguiente perdida de naturalidad (ver el capítulo 6 donde se describe todo con mayor nivel de detalle). El audio captado por el robot se usa fundamentalmente para las siguientes tareas: reconocimiento de voz basado en gramáticas, reconocimiento de voz de texto libre, detección de emociones en la voz, localización espacial de usuarios, cálculo de nivel de arousal (excitación) del entorno, generación de sonidos y acompañamiento musical que casen perfectamente con la voz recibida.

de investigación se han realizado la mayor parte de las pruebas, y consta de todos los canales de entrada y salida aquí enumerados

[7] http://en.wikipedia.org/wiki/Bluetooth

- Visión. El robot puede estar dotado de tres mecanismos basados en visión de percibir el entorno físico que le rodea: mediante cámara web (Maggie la incorporada en su boca [8]), cámara de profundidad (Maggie, usa una cámara Kinect [9]) y telémetros laser (Maggie incorpora un láser situado encima de la base móvil [10]). Estos sensores están dedicados a tareas, algunas actualmente en desarrollo, de: navegación por el entorno, detección e identificación de usuarios, detección de gestos y poses, detección de objetos y lectura de texto escrito (OCR).

- RFID. El robot puede estar equipado de uno o varios lectores de etiquetas de radio frecuencia (RFID), concretamente el robot Maggie, consta de varios lectores incorporados en su cuerpo, uno en la cabeza y otro en la base, ambos de corto alcance (unos 20 cm. para leer una tarjeta), y otros dos en sus costados de mayor alcance (unos 100 cm. aproximadamente). La interacción por etiquetas está destinada fundamentalmente a tareas de identificación de objetos, como por ejemplo medicamentos, o incluso al control del robot por diálogos por etiquetas. En este modo por etiquetas de radio frecuencia, cada una de ellas representa, mediante un dibujo adecuado (pictograma), cada una de las posibles habilidades del robot[11]. Este tipo de interacción es adecuada para niños pequeños, personas muy mayores o entornos muy ruidosos donde la interacción por voz se ve notablemente afectada.

- Tacto. El robot puede estar dotado de varios sensores capacitivos [12] capaces de detectar cuándo el usuario está tocando el robot en la parte del cuerpo donde esta situado dicho sensor. El sensor capacitivo no es capaz de notar diferencias de presión en el tacto, siendo únicamente binario (tocado / no tocado). Se usa como posible de entrada de información para el sistema de diálogo, así como para simular cosquillas en el robot.

- Smartphones[13] y tablets[14]. Es posible también la entrada de información mediante la tablet incorporada en el pecho del mismo o mediante teléfonos inteligentes. En ambos se presentan un conjunto de opciones (dependiendo del diálogo y la finalidad concreta de la interacción), que el usuario puede ir activando/desactivando mediante sus dedos.

[8] http://www.logitech.com/es-es/webcam-communications/webcams
[9] http://www.xbox.com/es-es/kinect
[10] $http://www.sick.com/group/EN/home/products/product_portfolio/optoelectronic_protective_devices/Pages/safety_las$
[11] El repertorio de habilidades del robot Maggie es muy variado. Cada habilidad tiene su tarjeta de radio frecuencia asociada
[12] http://en.wikipedia.org/wiki/Capacitivesensing
[13] http://en.wikipedia.org/wiki/Smartphone
[14] http://en.wikipedia.org/wiki/Tabletcomputer

3.2.2. Modos de salida de información

Como sistema de diálogo multimodal simétrico, la multimodalidad se presenta tanto a la entrada del sistema como a la salida del mismo, por lo que son necesarios diversos canales de salida o expresión de información:

- Audio. El robot Maggie tiene un sistema de cuatro altavoces situados debajo de su cabeza, los cuales permiten comunicarse con el usuario mediante voz y sonidos. Estos altavoces son usados para la generación de voz con emociones, sonidos no verbales, generación musical y reproducción de música.
- Gestos expresivos del diálogo (o como se conoce en inglés, "engagement gestures"). Mediante brazos, cabeza, párpados y base móvil el robot es capaz de realizar gestos que complementan el diálogo. Dentro de este repertorio de gestos se encuentran algunos como: negaciones, afirmaciones, seguir con la mirada al usuario, baile, navegación por el entorno. Estos gestos se realizan de manera sincronizada con la voz y los sonidos mediante el propio gestor de diálogo.
- Infrarrojo. Maggie es capaz de controlar electrodomésticos mediante comunicación por infrarrojos. En este sentido, el robot incorpora un mando de infrarrojos programable que permite emitir la señal adecuada para encender/apagar televisores, aires acondicionados, cadenas musicales, etc.

3.3. Componentes de RDS

El sistema de diálogo propuesto para nuestro robot (ver Fig. 3.3), necesita una entrada de información coherente en el tiempo y de contenido semántico, obtenida de fusionar la información que proporcionan cada uno de las entradas sensoriales posibles.

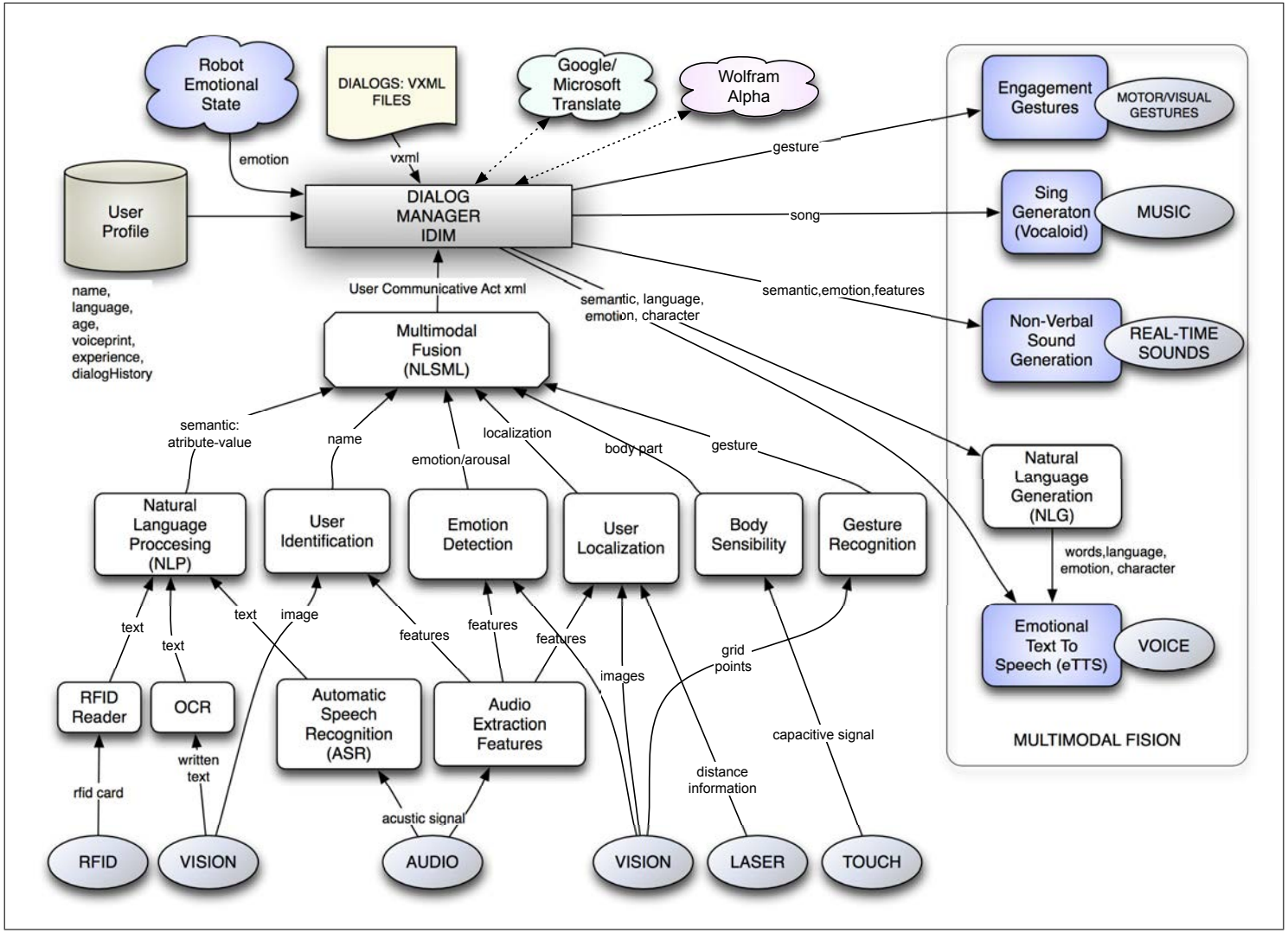


Figura 3.3: Esquema general del sistema de interacción RDS. A la derecha se representan los actuadores, en la parte inferior los sensores, y en la parte central el gestor del diálogo. Entre los sensores y el gestor de diálogo se ilustran varios módulos que procesan la información sensorial para entregarla al gestor de diálogo. En la parte derecha se encuentran los módulos relacionados con la expresividad del robot. El esquema general del sistema RDS actualmente esta siendo mejorado con la inclusión de un nuevo módulo de "fisión multimodal" que actúe de intermediario entre el gestor del diálogo y los módulos expresivos

.

El sistema de diálogo está compuesto por diversos componentes que realizan tareas específicas de manera desacoplada y distribuida. Estos componentes se pueden agrupar en:

1. **Gestor del Diálogo (IDiM)**: se encarga de gestionar los turnos de comunicación y dadas unas entradas de información, generar unas salidas. Es un gestor basado en el paradigma de rellenado de huecos de información, concretamente basado en el estándar Voice XML 2.1[15] extendido con funciones multimodales. Otros autores, entre los que destacan [Bennett et al., 2002] Eberman, [Eberman et al., 2002] [Kibria & Hellström, 2007] [Lucas, 2000] [Niklfeld et al., 2001], también usan este sistema de interacción, si bien casi ninguno de los estudios se ha realizado sobre robots sociales, estando la mayoría de ellos centrados en entornos telefónicos o web.

   Este paradigma trata de rellenar un hueco de información antes de pasar al siguiente. Para rellenar cada uno de los huecos de información es posible hacerlo mediante cualquiera de los canales de comunicación posible: voz, tacto, gestos, etc. o la combinación de ellos en un mismo mensaje. Un ejemplo sencillo puede ser intentar reservar un vuelo, en el que los huecos de información podrían ser: hora de salida, hora de llegada, ciudad origen, ciudad destino y compañía área. Todos estos huecos se pueden ir rellenando mediante sucesivas preguntas y respuestas entre el usuario y el sistema, o mediante una única frase, como por ej: "quiero salir de Madrid a las 7 de la mañana y llegar a París a las 9 de la mañana con Raynair".

   Recientemente se esta trabajando en la integración de un nuevo gestor de diálogo, conocido como *Iwaki*. Este gestor continua usando huecos de información (*slots*), pero con la ventaja de que es capaz de planificar la interacción mediante el uso de pre y post condiciones.

2. **Módulo de fusión multimodal**: cuando se habla de *fusión multimodal* es importante resaltar que esta se puede hacer a dos niveles (ver Fig 3.4). En el más bajo se encuentra la fusión que se realiza a nivel de módulos de percepción. Veámoslo con un ejemplo, para la tarea de localizar una persona respecto al robot, suele ser necesario de varias entradas de información que son fusionadas para conseguir una mayor precisión en dicha localización (audio, visión y láser). Este tipo de fusión de información, al nivel más bajo, para obtener una información más relevante semánticamente para el diálogo es lo que se entiende como fusión multimodal de bajo nivel.

   Aquí se hace referencia a un tipo de fusión multimodal de un nivel de abstracción más alto, en el que toda la información semánticamente relevante para el diálogo

[15] http://www.w3.org/TR/voicexml21/

(que es suministrada por otros módulos) tiene que ser fusionada temporalmente y entregada al gestor del diálogo. En este sentido, la fusión multimodal de alto nivel tiene más que ver con aspectos temporales y de identificación de los actos comunicativos que con las diversas fuentes de entrada de información ([Falb et al., 2007] [Shimokawa & Sawaragi, 2001]).

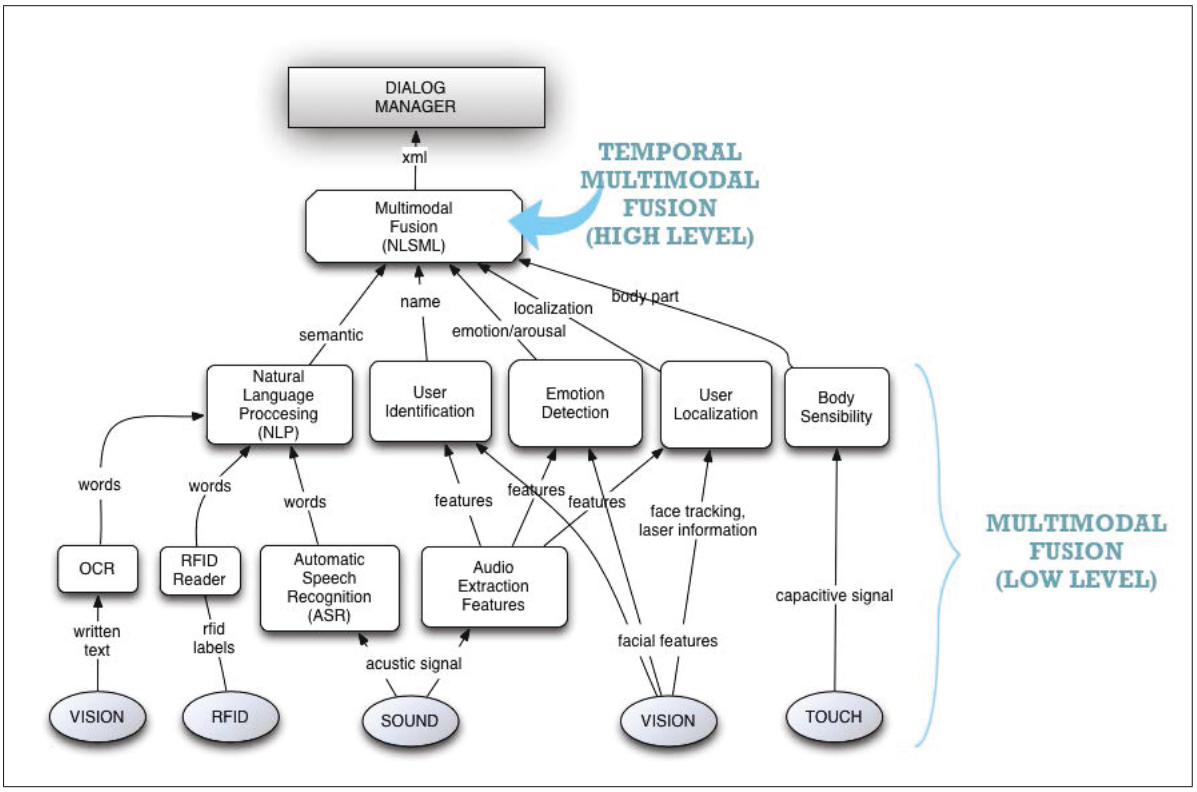


Figura 3.4: Niveles de Fusión Multimodal

Por lo tanto, la multimodalidad del sistema de diálogo nos brinda la posibilidad de comunicarnos con el sistema de diálogo por diferentes canales de comunicación: voz, gestos, tacto, tarjetas de radio frecuencia, sonidos, etc. La información o el mensaje transmitido simultáneamente por cada canal de comunicación conforma un mensaje comunicativo global a todos ellos. Por cada canal se transmite cierta información que complementa la transmitida por el resto de canales. Esta información que viaja por cada canal tiene cierto significado semánticamente relevante para el gestor del diálogo, y es generada por los distintos módulos del sistema de diálogo que trabajan directamente con los sensores del robot.

Es el módulo de fusión multimodal el que se encarga de agrupar la entrada de información recibida por cada uno de los canales en paquetes de información

mayores con un significado o mensaje global. En este mensaje global, que lo se conoce como *acto comunicativo*, el canal o los canales usados carecen de importancia. Por lo que, se puede afirmar que el módulo de fusión multimodal proporciona al gestor de diálogo la abstracción necesaria para obviar los canales usados para obtener la información relevante para el diálogo. Es la información recibida en los actos comunicativos la que sirve al gestor de diálogo para rellenar los huecos de información planteados en cada diálogo concreto.

3. **Módulos de entrada de información**: los componentes que trabajan sobre las entradas sensoriales para procesar dicha información y entregar al Gestor de Diálogo (IDiM) información relevante. Son los siguientes:

   - *Reconocimiento automático del habla (ASR)*: Se encarga de traducir voz a texto. Para ello, se ha implementado un sistema modular que permite conectar varios reconocedores de voz, que pueden trabajar en paralelo sobre las mismas muestras de audio. Actualmente se trabaja con el reconocedor de voz basado en gramáticas de *Loquendo* y el reconocedor basado en el servicio web de *Google ASR*.
   - *Procesamiento del lenguaje natural (NLP)*: se encarga de extraer información relevante del mensaje recibido ("semantic extraction"), con independencia del canal por el que se ha recibido dicho mensaje (reconocimiento de voz, de texto escrito, radio frecuencia...). Esta información extraída se enriquece o complementa con *información del mundo* ("*semantic enhacement*"). Esta información del mundo suele almacenarse en las llamadas *bases de conocimiento*, por lo que el enriquecimiento semántico se lleva a cabo mediante consultas a estas enormes bases de conocimiento.

     El tipo de información extraída suele corresponder a: pares atributo-valor, entidades, conceptos, la temática del discurso, "*sentiment*", expresiones temporales, dirección de correo, etc. Existen numerosos desarrollos que facilitan de manera comercial la extracción de esta información como son: *Semantria*[16], *Bitext*[17], *TextAlytics*[18], etc.

     Las bases de conocimiento más usadas son *Google FreeBase*[19], *Wolfram Alpha*[20], *Classora*[21], etc.

[16] `https://semantria.com/`
[17] `http://www.bitext.com/es/`
[18] `https://textalytics.com/inicio`
[19] `http://www.freebase.com/`
[20] `http://www.wolframalpha.com/`
[21] `http://www.classora.com/`

- *Conversor de texto escrito a texto máquina (OCR)*: reconoce y digitaliza el texto escrito a mano, para ello se usan técnicas de OCR. Para mayor detalle ver [Alonso-Martin et al., 2011].
- *Lectura de etiquetas de radio frecuencia (RFID)*: lee la información escrita en etiquetas de radio frecuencia tanto activas como pasivas. Para ampliar información consultar [A.Corrales, R.Rivas, 2009].
- *Identificación de usuarios*: este componente se encarga de identificar al usuario con el que está dialogando por su tono de voz. Para ello, en la fase de registro del usuario con el sistema, el sistema guarda información del timbre del usuario (voiceprints).
- *Identificación de emociones en los usuarios*: las emociones del usuario son captadas en función del tono de voz percibido por el sistema. Para ello, es necesario construir un clasificador que determine que emoción se corresponde, para dicho usuario, con las características sonoras extraídas de su voz, las *voiceprints* y la historia previa con el diálogo. Para ampliar consultar la publicación [Alonso-Martin et al., 2013b].
- *Localización de usuarios*: este componente se encarga mediante el sistema auditivo del robot, formado por 8 micrófonos y diferencias en amplitud de la señal recibida por cada uno de ellos, de localizar el origen de la fuente sonora. La localización sonora se apoya en la localización mediante láser para determinar con mayor precisión dónde se encuentra el usuario. Con esta información de disposición espacial, más la información que se han obtenido previamente en experimentos proxémicos del robot con usuarios reales, el sistema de diálogo puede determinar la situación más adecuada del robot frente al usuario. Para profundizar se recomienda leer la publicación [Alonso-Martín et al., 2012]. Recientemente se esta ampliando este sistema por uno más avanzado y preciso que incluye la información de profundidad suministrada por una cámara 3D.
- *Sistema del tacto*: este componente es capaz de detectar cuándo una extremidad del robot ha sido tocada. Actualmente no es capaz de determinar la presión ejercida por el usuario ni exactamente el gesto con el que ha sido tocado.
- *Detección de poses del usuario*: este componente se encarga de determinar la pose, es decir de clasificar la posición del cuerpo del usuario de entre las posibles. Es capaz de determinar si una persona está sentada, de pie, señalando a la izquierda, derecha o al frente, entre otros. Este sistema usa una cámara estereoscópica de profundidad y mecanismos de aprendizaje automático.

- *Otros*: existen componentes que se encargan de facilitar la conexión de entrada de información por tablet, smartphone o joypad.

4. **Módulos de salida de información**: los componentes que se encargan de expresar o transmitir el mensaje suministrado por el gestor del diálogo (IDiM) al usuario y que trabajan con las salidas del sistema. Ver la figura 3.3. Estos componentes son:

   - *Sistema de expresión verbal y sonora*: el módulo de eTTS (texto a voz con emociones) permite realizar tareas complejas, como son: gestionar la cola de luciones, traducir textos entre mas de 40 idiomas, además de adoptar distintos tonos de voz en función de los motores usados (*Loquendo, Festival, Microsoft TTS y Google TTS*). El módulo de "Non-Verbal Sound Generation" permite sintetizar sonidos no verbales en tiempo real, que permiten expresar al robots mensajes de manera similar a los producidos por la voz, pero en su propio lenguaje "robótico". Este tipo de sonidos se generan mediante el lenguaje de programación musical *Chuck*[22]. "*Sing Generation*" permite que el robot sea capaz de cantar mediante el software musical *Vocaloid*[23].
   - *Generación de lenguaje natural (NLG)*: Este sistema de se encarga de convertir información codificada en la computadora a lenguaje natural entendible por las personas. En la práctica se utiliza para construir frases en tiempo real partiendo de una idea. De esta manera, se consiguen diálogos más variados y naturales. Por ejemplo si la idea a transmitir es la de "saludar", puede ser convertido a los siguientes textos: "Hola, estoy encantado de hablar contigo", o bien "Hola, amigo", entre otros. Este mecanismo de convertir valores "semánticos" a "texto enlatado" puede hacerse de diversas formas. En nuestro sistema se ha usado un modelo muy sencillo de gramáticas o plantillas, muy similar a los usados en el reconocimiento de voz.
   - *Gestos expresivos*: permite al robot mediante sus extremidades (brazos, parpados, cabeza, cuello y base) realizar gestos típicos de cualquier diálogo, como son, negaciones, asentimientos, exclamaciones, incluso pasos de baile. Finalmente y aunque no es un componente propiamente dicho, el robot es capaz también de comunicarse mediante imágenes proyectadas en el tablet-PC, controlar componentes electrónicos mediante un mando infrarrojo, reproducir música descargada bajo demanda de Internet.

[22] `http://chuck.cs.princeton.edu/`
[23] `http://www.vocaloid.com/en/`

5. **Otros módulos**: además existen ciertos componentes presentes en el sistema de diálogo, que comunican con servicios web, que suministran entrada de información necesaria para el gestor de diálogo. Estos servicios usados son traductores automáticos (*Google Translate*[24] y *Microsoft Translate*[25]), buscador semántico (*Wolfram Alpha*[26]) y reproductor musical *Goear*([27]).

3.4. Principales características de RDS

Las principales características que describen al sistema aquí presentado son las siguientes:

- *Interpretado.* El sistema es interpretado, es decir se desacopla la especificación del diálogo concreto a cada situación (en ficheros de texto plano XML) de su interpretación y ejecución por el gestor del diálogo. Por lo tanto, se puede hablar de dos partes: la parte software, que es el propio Gestor de Diálogo IDiM, que ejecuta y interpreta los diálogos propiamente dichos. Y por otra parte, los diálogos quedan especificados en ficheros XML[28] que establecen ciertos huecos de información a rellenar.

- *Adaptable.* El diálogo puede adaptarse a cada usuario en base a características estáticas y dinámicas. Las estáticas son almacenadas en perfiles de usuario, que son aprendidas durante la interacción mediante diálogo natural con el usuario: idioma, nombre, edad y huellas de voz. Los factores dinámicos del usuario como son la experiencia con el sistema, la emoción detectada (computación afectiva), la situación espacial respecto al robot (proxémica) también sirven para personalizar la interacción. Los factores dinámicos del propio robot, como es su estado emocional también puede ser tenido en cuenta por cada diálogo específico. Esta adaptabilidad también se refleja en el multilingüismo: el diálogo es capaz de llevarse a cabo en multitud de idiomas.

- *Simétrico multimodal.* La interacción puede ser llevada por varios canales de entrada y salida de información. En este sentido, la multimodalidad se tiene en cuenta tanto a la entrada del sistema (fusión multimodal) como en las salidas (fisión multimodal). Los componentes del sistema de diálogo que lo convierten en simétrico multimodal han sido mencionados previamente.

[24] `http://translate.google.es/`
[25] `http://www.microsofttranslator.com/`
[26] http://www.wolframalpha.com/
[27] `http://www.goear.com/`,`http://www.splitcc.net/Downloads/playgoear.pl`
[28] http://en.wikipedia.org/wiki/VoiceXML

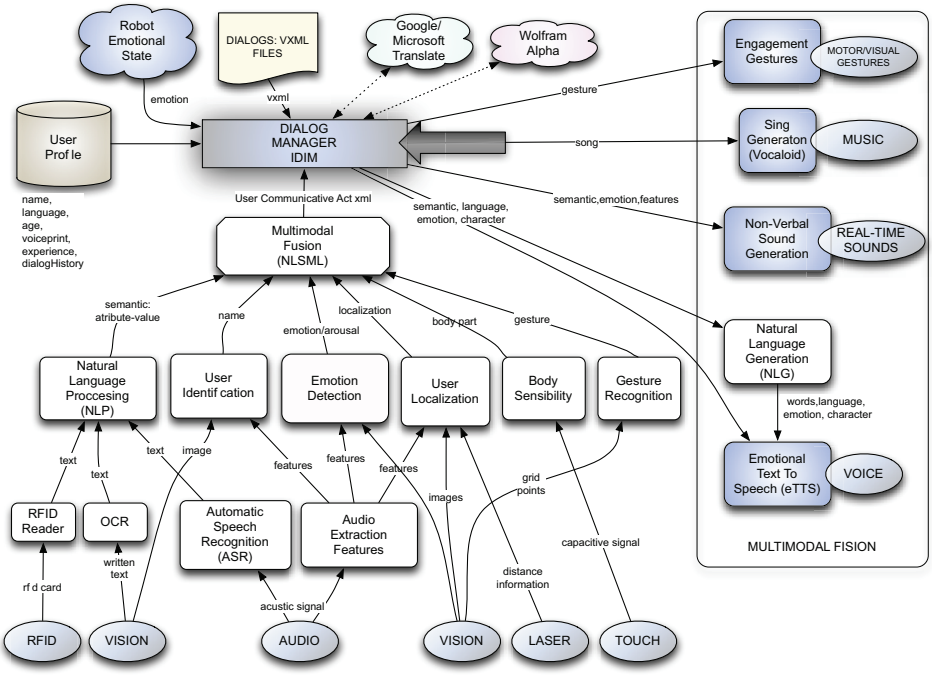
- *Multisonido*. Por multisonido se entiende un sistema en el que la entrada y salida sonora no se limita única y exclusivamente al reconocimiento y síntesis de voz. La voz puede ser usada para más tareas, además de existir otros medios sonoros no basados en voz. Siguiendo esta definición, el sistema multisonido aquí propuesto es capaz de analizar el audio de entrada para: realizar localización sonora, clasificar la voz del usuario en emociones, detectar el nivel de excitación-arousal del entorno, reconocer la voz, identificar al usuario. Es capaz de sintetizar sonidos para: generar voz con emociones, generar sonidos no verbales robóticos en tiempo real que aumentan notablemente la expresividad del robot, expresarse musicalmente mediante cante y finalmente la capacidad de reproducir de música online.

3.5. Resumen

En este capítulo se presentó el sistema de interacción natural entre robots y humanos llamado RDS. Este sistema convive con el resto de sistemas que forman parte del sistema operativo que controla cada robot, AD-ROS. Se ha realizado una descripción general del sistema y de los componentes que lo forman, los canales/modos que gestiona el sistema así como sus principales características. Una descripción pormenorizada de cada componente se hace en los siguientes capítulos.

CAPÍTULO 4

Gestión del diálogo



"Todo el mundo en este país debería aprender a programar un ordenador, porque eso te enseña a pensar."— Steve Jobs

4.1. Introducción

En el capítulo anterior se ha presentado los rasgos generales del sistema de diálogo RDS en su conjunto, haciendo una descripción por módulos y funcionalidad. Este capítulo se centra en describir el componente principal del sistema, ya que es el que se encarga de la gestión del diálogo propiamente dicho.

Los gestores de diálogo implementados hasta la fecha en robots sociales suelen estar fuertemente acoplados a una habilidad concreta del robot cuyo fin es bastante específico. Esto supone una serie de limitaciones claras, tanto si se quieren realizar modificaciones en el diálogo, como si el conjunto de habilidades del robot es amplio, dinámico y variable. En este trabajo se presenta un nuevo gestor de diálogo, al que se ha denominado IDiM, que separa el manejador de diálogo de la implementación del diálogo en sí misma. De este modo, el manejador de diálogo carga, interpreta y navega entre distintas implementaciones de diálogos en tiempo de ejecución. El gestor presentado consigue unir las necesidades semánticas de cada habilidad interactiva del robot, con una implementación de diálogo para cada habilidad. Se forma así un mapa semántico o un contexto de diálogo concreto que se va modificándose de acuerdo a la ejecución secuencial de cada una de sus habilidades.

En este capítulo se muestra el diseño e implementación del manejador de sistema de diálogo dinámico, IDiM. La principal contribución del sistema es que es capaz de cambiar el contexto del diálogo dinámicamente, esto es, sin hacer ningún cambio en el sistema propiamente dicho. Otra ventaja de IDiM es que es capaz de manejar diferentes e ilimitadas interfaces multimodales debido a su modo estandarizado de representar la información relevante por todos los módulos de percepción. Finalmente, y como ejemplo de su uso se presenta IDiM como un integrador de habilidades[1], siendo capaz de arrancar y parar mediante interacción natural cualquiera de ellas.

El hecho de que diferentes implementaciones de diálogos puedan ser cargados y descargados en tiempo de ejecución tiene varias ventajas: sólo hay un gestor del diálogo (DMS) para cualquier número de diálogos; es posible asociar diferentes diálogos para cada habilidad interactiva; el sistema hace sencilla la tarea de implementar nuevos diálogos, dado que es interpretado en tiempo de ejecución, puede ser modificado sin necesidad de modificar el código fuente de la habilidad. El código fuente de la habilidad permanece mientras que el diálogo se especifica mediante un sencillo lenguaje

[1]En el siguiente capítulo se mostrará detalladamente un diálogo implementado en el que el gestor se encarga de ir activando o desactivando habilidades del robot. Por habilidad se entiende una capacidad del robot para desempeñar una cierta tarea. El repertorio de habilidades del que consta el robot Maggie es bastante amplio. Algunas de estas habilidades son puramente sensoriales, mientras que otras son puramente motoras, otras en cambio son una combinación de ambos tipos. El nivel de abstracción y complejidad de cada una de estas habilidades es muy variado, siendo muy común la cooperación entre ellas para lograr tareas más complejas

en un fichero independiente, que no es necesario recompilar [2].

4.2. Gestores de diálogo en robots sociales

Esta sección corresponde con un estado del arte de los gestores de diálogo en la interacción humano-robot atendiendo a lo que se ha considerado las características más importantes extraídas del análisis de la literatura: estructuras de gestión del diálogo, expresividad y multimodalidad, representación del conocimiento, conocimiento compartido, comunicación cooperativa y coordinada, diálogos y aprendizaje, conversación multiparte, escenarios y métricas para una evaluación apropiada del sistema.

Hay algunas aproximaciones teóricas de alto nivel sobre un gestor de diálogo; también hay implementaciones de gestores de diálogo de alto nivel para agentes virtuales. Sin embargo, y poniendo énfasis en la importancia de un cuerpo físico y robótico (*embodiment*) y cómo esta característica afecta a la interacción natural, este trabajo se ha focalizado en la implementación y evaluación de estos sistemas en robots reales.

4.2.1. Estructuras para la gestión del diálogo: arquitecturas y escalabilidad

La gestión del diálogo ha sido ampliamente estudiada en sistemas de inteligencia artificial (agentes y agentes virtuales). Analizando tales sistemas se pueden considerar dos aproximaciones diferentes: gestores que están basados en patrones, y no tienen en cuenta el significado semántico de las oraciones, como por ejemplo *ELIZA* ([Weizenbaum, 1966]), *AIML* o *ALICE* ([Wallace, 2000]); y gestores que trabajan usando algún tipo de razonamiento semántico o teoría de diálogo natural, como por ejemplo *TRINDI* [Larsson et al., 2004], o, enfocándonos directamente a la robótica, *RavenClaw* ([Bohus & Rudnicky, 2009]).

Los gestores del diálogo *(Dialog Management System*: DMS) se encargan de mantener y actualizar la información del contexto del diálogo así como decidir la siguiente acción a tomar, o el siguiente diálogo a ejecutar. En ese sentido, actúa como interfaz entre los diversos componentes de la arquitectura de control y el usuario.

EL DMS elige que *acto comunicativo* (AC) debe ser interpretado en un estado concreto del diálogo. Normalmente, los posibles AC se encuentran codificados de antemano por el desarrollador del sistema: en cada estado del diálogo, el desarrollador ha codificado un conjunto posible de tipos de acciones y es el *DMS* el que elige la apropiada de entre el conjunto de posibles.

La interacción natural responde a algunos patrones estándar de comportamiento ([Schegloff & Sacks, 1973]). Algunos *DMS* aplicados a robótica están basados en esos

[2] http://es.wikipedia.org/wiki/Compilador

patrones de interacción. Por ejemplo *PAMINI DMS* ([Peltason & Wrede, 2010b]), donde esos patrones son independientes de la tarea del diálogo para la que fueron concebidos.

En [Henderson et al., 2008] se usa un método híbrido, que combina como estrategias de gestión de diálogo ‘‘*aprendizaje por refuerzo*”[3] con *“aprendizaje supervisado*”[4]. El proceso de aprendizaje es hecho usando gran cantidad de ejemplos de diálogos de un corpus. Este método fue aplicado en el sistema de diálogo *COMUNICATOR*.

Aproximaciones similares usando aprendizaje por refuerzo son encontradas en [Goddeau & Pineau,], donde un método de aprendizaje por refuerzo fue aplicado al sistema *DIALER* de *AT&T*. El *DMS* seleccionaba entre una gran cantidad de posibles *actos comunicativos* para alcanzar un objetivo concreto. Por ejemplo, obtener un número de teléfono de una larga lista telefónica haciendo preguntas al usuario.

En otro trabajo, el estado del diálogo se corresponde con las intenciones del usuario. Por ejemplo en [Roy et al., 2000] usando un pseudo POMDP (Parcialmente Observable Proceso de Decisión de Markov), el *DMS* es capaz de compensar la alta incertidumbre del proceso de reconocimiento de voz (baja precisión), en base a mantener “n” hipótesis del estado del diálogo simultáneamente, basándose en los resultados de reconocimiento de voz obtenidos. Las acciones son las respuestas llevadas a cabo por el sistema, de las cuales se tienen 20 diferentes (10 se corresponden a habilidades del robot y las otras 10 para tareas de aclaración o confirmación). El *POMDP* planifica sobre un espacio de creencias; cada creencia consiste en una distribución de probabilidades sobre el conjunto de estados, representando la probabilidad de que el diálogo se encuentre en ese determinado estado. Las observaciones son obtenidas a través del reconocimiento de voz y como se ha comentados las acciones son las respuestas que el sistema lleva a cabo (en este caso 20 posibles).

Existen otros gestores de diálogo basados en *agentes inteligentes* que colaboran para llevar a cabo una determinada tarea. Son muy adecuados para tareas complejas. Finalmente, se puede traducir la interacción entre agentes como un grafo o árbol dinámico con reglas: precondiciones-postcondiciones, por ello también son conocidos como sistemas basados en reglas.

Ravenclaw/Olympus [Bohus et al., 2007] o JASPIS [Turunen et al., 2005] se basan

[3]Programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento.

[4]Es una técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento consisten de pares de objetos (normalmente vectores): una componente del par son los datos de entrada y el otro, los resultados deseados. La salida de la función puede ser un valor numérico (como en los problemas de regresión) o una etiqueta de clase (como en los de clasificación). El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos, los datos de entrenamiento. Para ello, tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente.

en arquitecturas con varios *agentes inteligentes*, que consisten en módulos independientes trabajando en paralelo para lograr una cierta tarea. Olympus está basado en el *CMU Communicator/Galaxy* [Seneff et al., 1996]. Cada módulo está conectado a los otros mediante un nodo central (hub). El *DMS* RavenClaw representa el dialogo como un conjunto de árbol de nodos con sus respectivas pre y post condiciones.

4.2.2. Expresividad y multimodalidad

La expresividad esta relacionada con cómo el robot puede expresarse usando todas sus modalidades, esto es, cómo afronta el problema de la "fisión multimodal". Tradicionalmente la multimodalidad ha estado más relacionada con el problema de "fusión", es decir, de cómo unificar o agrupar la información recibida por diferentes módulos en un mensaje comunicativo coherente con información relevante para el gestor del diálogo.

Desde un punto de vista verbal, las oraciones expresadas verbalmente son muy diferentes de las escritas. Una oración verbal, incluye o puede incluir falsos comienzos, dudas, repeticiones, frases cortadas o incompletas, y otras muchas variantes a lo largo del tiempo. Estos factores dificultan el reconocimiento de voz, por ser componentes básicos en la expresividad de lenguaje natural, dado que además de transmitir información puramente verbal, también transmiten cierta información no verbal, como velocidad y ritmo comunicativo, características prosódicas como el timbre y frecuencia de la voz, volumen, etc. que sirven para tareas como: identificación del usuario, género, edad, intención, humor, aptitud, emoción, etc.

Por otro lado, el discurso es normalmente acompañado de expresiones no verbales como acciones kinésicas[5], indicadores visuales, aspecto físico del interlocutor, etc. Por lo tanto, aparte del discurso, un robot que quiera llevar a cabo una interacción natural por voz debería manejar estas modalidades no verbales de percepción y expresión para reconocer/expresar gestos, disfluencias prosódicas[6], mensajes audioviusales, etc. de una manera sincronizada.

Los gestos en un robot normalmente son implementados usando un

[5]La kinésica también es conocida con el nombre de comportamiento kinésico o lenguaje corporal. También puede definirse como el termino para las formas de comunicación en los que se intervienen movimientos corporales y gestos, en vez de (o además de) los sonidos, el lenguaje verbal u otras formas de comunicación.

[6]Los titubeos, las repeticiones, los falsos comienzos y las pausas representan disfluencias lingüísticas. Los hablantes no tendemos a percibir la presencia de estas interrupciones en el ritmo continuo del habla espontánea, a pesar de su frecuente ocurrencia. Se ha demostrado que las vacilaciones o pausas vacías surgen normalmente tras un intervalo de siete u ocho palabras durante el intercambio lingüístico. Para ampliar consultar: `http://www.alonsoquijano.org/cursos2004/animateca/recursos/Biblioteca%20virtual/Lectuario%202002/Febrero-Estudios%20de%20linguistica/Eloy%20&%20Juan%20de%20Dios.htm`

diccionario de gestos, llama *gestionario*. Por ejemplo, los gestos del robot *Leonardo* ([Breazeal et al., 2004]) son basados en un gestionario; *Robovie* ([Ishiguro et al., 2002]) integra un gestionario dentro de los propios "prompts" a sintetizar por el *TTS* (menajes "enlatados" con el texto a sintetizar por la herramienta de texto a voz en los que se entremezclan gestos del diálogo). Otras aproximaciones, no basadas en gestionarios, tratan de generar gestos icónicos sin dar una información explícita del gesto a realizar.

En los sistemas mostrados en [Cassimatis et al., 2004], [Perzanowski et al., 2001b], [Lemon et al., 2002], y otros muchos trabajos similares, la información verbal es fusionada con gestos táctiles sobre una pantalla capacitiva, que muestra algún tipo de información, como mapas y menús. Este tipo de sistemas son capaces de interpretar oraciones como "ve allí" (mientras se señala un punto en el mapa de la pantalla). En [Lemon et al., 2002] la interfaz visual es también usada por el *DMS* para mostrar vídeos mientras se genera y sintetiza el lenguaje verbal. Por ejemplo, preguntando "¿es ese el coche rojo que estás buscando?" (mientras se muestra un vídeo de lo que la cámara del robot detecta). En [Hüwel et al., 2006] y [Toptsis et al., 2004] el robot BIRON es presentado. Su sistema de diálogo fusiona conjuntamente información visual con información verbal (deíxis), por lo que es capaz de resolver oraciones como "eso es una planta (mientras señala una planta)".

La *fusión multimodal* ha sido usada en algunos trabajos de "sistemas de programación natural". Por ejemplo, en [Iba et al., 2002] un robot aspiradora es programado usando gestos y voz, con lo que el robot es capaz de fusionar dos canales de información e inferir acciones que el usuario quiere llevar a cabo. Se encuentran trabajos similares en [Dominey et al., 2007], donde el robot HRP-2 fusiona información visual y verbal para manipulación de objetos.

El gestor de diálogo *TRINDI* ([Larsson et al., 2004]) ha sido usado en escenarios e interacción humano-robot multimodales, donde ambos colaboraban para resolver una tarea que requería diálogo en ([Burke et al., 2002]). Trabajos similares se han realizado con agentes virtuales, usando el gestor *DIPPER*, una secuela de *TRINDI* ([Bos & Oka, 2003]).

4.2.3. Representación del conocimiento: modelo del mundo y del usuario

Hablar con robots implica que el robot implementa o debería implementar un modelo del mundo, sobre el que el sistema pueda cargar y grabar información sobre el humano y el entorno con los que interactúa. Cómo representar ese modelo supone un problema realmente importante, sin embargo ese modelo del mundo tiene que estar unido al propio diálogo.

Una de las primeras organizaciones de diferentes elementos del discurso fue la usada en [Grosz & Sidner, 1986]. El contexto del diálogo, fue dividido en tres componentes diferentes: uno para la estructura de la secuencia de oraciones, otro para la estructura del propósito o intención embebida en dicha oración, y otro componente para el estado atencional, es decir, un registro "de los objetos, propiedades y relaciones que son de interés en cada punto del discurso".

Un ejemplo robótico de un gestor que usa un modelo del entorno puede ser encontrado en [Kruijff et al., 2007]. En él, un robot móvil aprende características del entorno. El sistema descrito representa como el entorno es estructurado usando ontologías. Las representaciones espaciales son realizadas en tres capas: mapa de métricas, con información geométrica o de odometría[7], para la navegación del robot; mapa topológico, que divide el grafo de navegación en áreas delimitadas por "nodos puerta"; y finalmente el mapa conceptual que guarda nombres de áreas y objetos presentes en ellas, y permite inferencias como que un "concepto" (habitación o objeto) pueda ser verbalmente referenciado. La representación está unida con el diálogo mediante una "'*Combinatorial Categorical Grammar* (CCG)"' que representa oraciones (tanto de reconocimiento como de generación de lenguaje) desde un punto de vista semántico.

En el sistema TRINDI ([Larsson et al., 2004]) o en *DIPPER* ([Bos et al., 2003]) el sistema de conocimiento es implementado con "*estados de información*" (EI). Básicamente cada EI esta formado por un conjunto de huecos de información asociados con unas reglas de actualización (escritas en el lenguaje de programación lógica PROLOG) con la siguiente estructura: *<nombre, condiciones, efectos>*. Esas reglas definen cómo obtener la información. El principal objetivo del sistema es rellenar esos huecos de información para cada EI. Cuando hay algún cambio en dichos EI, se dice que hay un "movimiento en el diálogo" ([Traum, 2000]).

Los llamados sistemas basados en formularios, como cabía esperar por su nombre, representan el diálogo como formularios anidados compuestos a su vez por huecos de información a rellenar. El estándar W3C *VoiceXML* [8] implementa este tipo de paradigmas y ha sido llevado al campo de la robótica con éxito en sistemas tales como [Niklfeld et al., 2001], [Nyberg et al., 2002], [Bennett et al., 2002] o en [Schmitz et al., 2008] donde el robot humanoide ROMAN es descrito. Este robot social es capaz de buscar y encontrar gente en el entorno. El principal objetivo es dialogar con ellos de manera natural. Su *DMS* está basado en una descripción XML muy similar a la realizada por *VoiceXML*. Nuestro sistema, como se verá posteriormente a lo largo del capítulo, se corresponde con un sistema basado en formularios y huecos de información.

[7]La odometría es el estudio de la estimación de la posición de vehículos con ruedas durante la navegación. Para realizar esta estimación se usa información sobre la rotación de las ruedas para estimar cambios en la posición a lo largo del tiempo

[8]http://en.wikipedia.org/wiki/VoiceXML

En los últimos tiempos, el modelo del mundo se está centrando, en lo que concierne a los gestores de diálogo, en los usuarios. En este sentido, en sistemas como [López-Cózar & Callejas, 2006] desarrollado en la Universidad de Granada (España), y nuestro propio sistema aquí presentado, mantiene un "perfil de usuario" de cada uno de los usuarios conocidos en la interacción con el sistema. Este perfil de usuario es almacenado en una base de datos, en la que se mantiene información relativa al usuario, como son: el nombre, la edad, la experiencia, el idioma preferido, las características de su timbre de voz, el esquema de colores de la vestimenta, la experiencia total de uso con el sistema, la historia previa del diálogo, determinada información proxémica, etc. Todos estos factores constituyen información muy valiosa para la personalización/adaptación de los diálogos a cada usuario concreto.

4.2.4. Conocimiento compartido

Los conceptos de "conocimiento compartido (*Common Ground*)" ([Clark, 1996]) y "campo de experiencia (*Field of Experience*)" ([Schramm, 1954]) son introducidos en sendos trabajos. En diálogos naturales, el "conocimiento" es obtenido mediante interacción: una interacción da una idea de alguna cosa y el otro/s interlocutor/es da/n su *feedback* positivo o negativo (realimentación positiva o negativa), o solicita dicha realimentación. *Feedback* positivo podría ser solo una confirmación implícita o explícita. En cambio, *feedback* negativo podría indicar una falta de entendimiento del mensaje, una petición para repetir la oración o una petición de aclaración.

Se puede definir el concepto de "*Grounding*" como el proceso en el que el robot realiza la unión de cierta abstracción de sus medidas sensoriales, relacionados con un elemento delimitado externo, o de cierto acto dentro de sus habilidades, para el cual se asigna una etiqueta o referencia en el lenguaje. Por ejemplo, si el robot es capaz de encender una televisión externa mediante un actuador por *IR*, existirá *grounding*, cuando sea capaz de realizar la unión entre los *actos comunicativos* que referencian el acto de "encender la tele" con el propio acto de encenderla.

Notar el papel importante del silencio en la interacción natural. La mayoría de los *DMS* actualizan el estado del diálogo cuando reciben una oración por parte del usuario, es decir cuando detectan cierto tiempo de silencio ([Roy et al., 2000] y muchos otros); pero el silencio también puede ser visto como un *acto comunicativo* propiamente dicho, que debería ser manejado, con significados semánticos adecuados, como por ejemplo que el usuario no ha entendido, o que no escucha, que esta dubitativo y no sabe que acción llevar a cabo, que se ha "desenganchado", etc.

Los modelos de "conocimiento" en humanos constan de varios niveles, desde el conocimiento compartido sobre el mundo físico (leyes física, reglas para manipular objetos, ciertos conocimientos específicos, etc) hasta convenciones sociales y reglas culturales sobre como comunicarse. Los humanos entienden o tratan de entender las

intenciones que se esconden detrás de cada *acto comunicativo*. A mayor nivel de "conocimiento compartido" entre los usuarios, menor es la cantidad de señales necesarias para la interacción. Un importante componente del "conocimiento" es el conocido como "mantener la atención (*Joint Attention*)' ([Tomasello, 2008]) o el "reconocimiento de la intención" ([Allen et al., 2001]). El "mantener la atención" es la capacidad de los interlocutores de coincidir en el interés por el asunto (físico o conceptual) por el que se está dialogando. El reconocimiento de la intención implica que los usuarios reconocen planes y objetivos del otro interlocutor, gracias a la percepción e interpretación de sus *actos comunicativos*.

Dado que el conocimiento compartido ayuda a hacer más efectivo el diálogo, también es tenido en cuenta en robótica social, como es mostrado en [Iwahashi et al., 2010]. El robot *CoBot* ayuda a los usuarios en reuniones dando información relevante para las peticiones de los usuarios y también tomando la iniciativa si el robot llega a no estar seguro de su estado actual. El robot pasa un día entero con el usuario (interacción a largo plazo). Durante ese día, el sistema intenta reducir la repetición de diálogos, dando más detalles a cada usuario que pasa por la misma localización, usando sinónimos en el proceso de generación de lenguaje natural. Por lo tanto, el sistema añade nueva información sobre la experiencia compartida: lugares que ellos han ya visitado, que pasa a ser parte del conocimiento compartido entre el usuario y el robot.

"*Generalized Grounding Graphs* (G^3)" es un entorno de desarrollo usado para relacionar palabras con aspectos del mundo exterior: objetos concretos, lugares, caminos y eventos ([Tellex et al., 2011]). Algo similar al trabajo que está desarrollando Google con el llamado "grafo del conocimiento", para hacer búsquedas semánticas más inteligentes, que una simple búsqueda en páginas webs [9]. La mayoría de los sistemas manualmente conectan entidades del lenguaje (básicamente expresiones regulares de palabras) sobre un delimitado espacio de acciones fijadas por el desarrollador. Por ejemplo, el robot *Florence Nightingale* (Flo), es un robot enfermero que realiza una relación entre las creencias y acciones del diálogo (normalmente acciones de hablar), que son especificadas, a priori, por el programador. G^3 también ha sido usado para tareas de manipulación, donde un robot manipulador móvil obedece a ordenes como "pon el brazo neumático sobre el camión". El robot lleva a cabo inferencias sobre este tipo de comandos en lenguaje natural para crear y ejecutar un plan correcto.

Un ejemplo de "mantener la atención" en robótica podría ser el encontrado en [Haasch et al., 2004]. *BIRON* es usado para estudiar como el robot es capaz de detectar y seguir el "asunto" expresado por el usuario. El robot interactúa relacionando objetos con palabras.

[9] http://www.fayerwayer.com/2012/05/google-presenta-el-grafo-del-conocimiento-para-darle-sentido-a-las-busquedas/

4.2.5. Comunicación cooperativa y coordinada

La comunicación natural, como debe ser tomada en el proceso de diálogo Humano-Robot, implica una adaptación recursiva entre el comunicador y el oyente ([Tomasello, 2008]). Por lo tanto, la comunicación puede ser vista como una actividad cooperativa. La cooperación tiene que estar bien coordinada a lo largo del tiempo (ritmos comunicativos, velocidades, etc) y tiene que ser coherente, desde un punto de vista semántico.

En interacción natural, cada turno es intercambiado asíncronamente durante el proceso de interacción. Ese intercambio de turno es normalmente establecido gracias a señales no verbales. Durante la interacción cada interlocutor puede tomar la iniciativa, esto en un DMS puede ser gestionado de tres formas diferentes:

- La iniciativa la lleva el sistema: el sistema pregunta para obtener cierta información, mientras que el usuario se limita a contestar a dichas preguntas. Por ejemplo, un sistema de huecos de información para por comprar un billete de tren.

- La iniciativa es del usuario: igual que el anterior, pero es el usuario el que hace las preguntas y el sistema responde. Ejemplos de este sistema, son aquellos que responden a todo tipo de preguntas en lenguaje natural (por ejemplo realizando consultas a buscadores semánticos) o robots que actúan como esclavos limitándose a ejecutar las ordenes que ordena "el amo" (el humano).

- De iniciativa mixta: ambos, el sistema o el usuario son capaces de iniciar e interrumpir la conversación, tomando la iniciativa bruscamente en el intercambio de turno, o solicitándola mediante lenguaje no verbal. Un ejemplo de este tipo de iniciativa es la que se puede observar en programas de debate, en el que los contertulios, constantemente toman la iniciativa, llegando incluso a cortar al interlocutor que se encontraba en el uso de la palabra.

La iniciativa mixta ha sido usada en robótica gracias a algunos DMS's ([Allen, 1999]). Por ejemplo, el robot móvil descrito en [Kruijff et al., 2006] es capaz de aprender cosas del entorno gracias a la interacción con el usuario, en modo iniciativa mixta; el robot puede preguntar por cierta información, y a su vez el usuario puede solicitar información sin petición previa del robot. También el sistema de diálogo *PAMINI* [Peltason & Wrede, 2010b] fue concebido para una interacción con iniciativa mixta, usando patrones de interacción.

La colaboración también esta presente a otros niveles, por ejemplo participando en una tarea común y compartiendo planes. En [Kidd et al., 2004], es presentada la idea de un robot humanoide como un socio colaborador del humano. Otro robot, bastante conocido en la literatura es el robot *Leonardo*, que también tiene ciertas capacidades

de interacción no verbales, para llamar y mantener la atención, como la mirada y señalando. Es capaz de aprender nombres de objetos desde *actos comunicativos* del usuario. Por ejemplo, para la oración "Lenardo, eso es *(gesto deíctico)* un botón azul", el robot incorpora la etiqueta lingüística "botón azul" a la información visual recibida sobre el objeto percibido. A su vez como se ha comentado, *Leonardo* tiene varios gestos no verbales, que ayudan en el aspecto colaborativo de la comunicación, como por ejemplo, si hay una falta de compresión, *Leonardo* lo puede expresar con un determinado gesto (por ejemplo arquear las cejas), o si el usuario comienza a hablar, el robot expresa mediante un determinado gesto, como puede ser una mirada, que está manteniendo la atención, etc. Por otro lado, tanto el robot como el usuario son capaces de colaborar en un mismo plan: *Leonardo* completa con una acción (por ejemplo pulsar un botón) una secuencia aprendida (plan) si el usuario no lo hace.

Otro robot humanoide colaborativo y de iniciativa mixta es *ROMAN* [Schmitz et al., 2008], dotado de un complejo sistemas de diálogo. El objetivo de ROMAN es lograr una interacción con humanos lo más natural posible en entornos públicos, para ello usa un gestor de diálogo multimodal, que trabaja con información verbal y no verbal. El comportamiento del humanoide es como sigue: el robot está mirando a su entorno buscando personas. Una vez que una persona es detectada, *ROMAN* trata de atraer la atención del humano hablándole. El humano puede mostrar interés respondiéndole. Si en cambio, no muestra interés el robot dice "es una lástima" y busca otras personas. En cambio, si el humano muestra interés, el robot se presenta y responde a las preguntas que el humano pueda hacer sobre el mismo y el grupo que lo desarrolla. Si el humano pierde el interés y se aleja del robot la conversación finaliza. Los diálogos están formalizados en documentos XML siguiendo el paradigma de VoiceXML.

En [Carnegie & Kiesler, 2002] es estudiado como un humano y un robot pueden cooperar en tareas diarias. El artículo esta enfocado sobre como los humanos crean modelos mentales sobre el comportamiento de un robot asistente con diferente personalidades.

4.2.6. Contextualización del diálogo

Los robots, como agentes conversacionales dotados de cuerpo físico, mientras se encuentran en un entorno dinámico, tienen que adaptar sus diálogos a cambios en el entorno, es decir, tienen que ser *contextualmente* apropiados. En este trabajo, se entiende la contextualización como el proceso de resolución de referencias entre un modelo del dominio del entorno y el discurso. Una apropiada contextualización implica conocimiento externo (modelo del mundo y del usuario) e interno del robot (propia historia previa del robot y de los diálogos realizados, información compartida por otros robots, etc). Por ejemplo, un sistema de diálogo contextualizado trata de

resolver oraciones como “gira a mi derecha”, “es la segunda habitación después del laboratorio”, “un poco más”, etc. Para un diálogo contextualizado, un modelo del entorno y del espacio de estados del usuario resolverían estas ordenes. Esos gestores de diálogo normalmente combinan sus modeles internos con el reconocimiento de voz y el dominio de acciones del robot.

Diálogos contextualizados son muy importantes para robots móviles cuyo objetivos principales son: adquirir o aprender como el entorno es semánticamente/conceptualmente estructurado ([Kruijff et al., 2007]), aprender rutas por las que moverse en una ciudad ([Bugmann et al.,]), obedecer una orden con algún tipo de deixis ([Trafton et al., 2006]), colaborar en una tarea donde el robot tiene que tener en cuenta el punto de vista del usuario ([Schultz, 2004], [Skubic et al., 2004] y muchos otros).

Por lo tanto, en un diálogo contextualizado, los interlocutores tienen que tratar con diferente información que proviene de diferentes modalidades que ayudan a que el sistema resuelva diferentes tipos de ambigüedades en el lenguaje (deixis).

4.2.7. Diálogo y aprendizaje

Una cuestión importante en aprendizaje es como comunicar la adquisición de nuevas palabras, y otro problema es el relacionado con que hacer cuando esas palabras tienen múltiples significados ([Chauhan & Lopes, 2010]). El sistema *LCore* combina información multimodal (voz, visual y táctil) para lograr que el robot aumente su espacio de creencias (modelo del entorno). En [Taguchi et al., 2009] un manipulador robótico es capaz de dialogar para entender el significado de ciertas palabras como “cual” o “que” gracias a un aprendizaje probabilístico. El sistema es capaz de aprender como unir oraciones (que involucran voz) con objetos en ciertos entornos muy restringidos (que involucran visión).

En un modelo de aprendizaje basado en instrucciones (IBL) o “Programación Natural de Robots (NRP)”, el robot aprende nuevas tareas a través de las instrucciones dadas por el propio usuario, sin necesidad de que el programador implemente nuevo código fuente para la ejecución de esas tareas o instrucciones. Este tipo de aprendizaje trata de asemejarse a la manera en que los humanos aprendemos a realizar nuevas tareas. Por ejemplo, TRINDI ha sido usado como un DMS (manejador de diálogo) basado en IBL (aprendizaje por instrucciones) donde un robot móvil aprende nuevos caminos en una maqueta de una ciudad ([Lauria et al., 2002],[Lauria et al., 2001]).

Nuestro *DMS*, llamado **IDiM**, descrito en este capítulo, ha sido también usado satisfactoriamente para tareas de *NRP*; permitiendo a través de diálogos naturales (fundamentalmente basados en voz y tacto en nuestro robot social Maggie), el aprendizaje de nuevas secuencias de movimientos, válidas para saludos, pasos de baile, etc ([Gorostiza & Salichs, 2011]).

Volviendo nuevamente al robot *BIRON* ([Spexard et al., 2006]), este es capaz de aprender el nombre de nuevos objetos y lugares del entorno a través de *fusión multimodal*, mientras conversa en interacción natural con el usuario en un ambiente doméstico. También el robot Leonardo, como se comento anteriormente, es capaz de aprender secuencias simples de acciones (como pulsar un botón de un determinado color) mediante diálogos, en ese caso, más basado en interacción mediante gestos ([Kidd et al., 2004]) que mediante voz.

4.2.8. Conversación multiparte (varios interlocutores simultáneos)

La mayoría de los *DMS* son diádicos, es decir, son concebidos para una interacción de solo dos partes: el robot y el usuario. Un *DMS* multiparte tiene que tener en cuenta un mundo abierto de interacciones y sin restricciones en el que involucran más de dos participantes de ambos tipos, humanos y robots. Desde un punto de vista externo, uno de los mayores problemas de tales sistemas es como distinguir quién tiene el turno de palabra, quién está hablando, quien está escuchando, etc. Desde un punto de vista interno, o de programación, para lograrlo se necesitan tareas adicionales de localización de usuarios y fuentes sonoras, separación del audio en un canal para cada interlocutor, carga de diferentes perfiles de usuario al mismo tiempo, mantener por cada interlocutor diferentes creencias, objetivos, deseos e intenciones ([Bohus & Horvitz, 2010]).

Un importante desafío es que el sistema dirija su capacidad de interrumpir una interacción con alguien, para retomarla posteriormente en un futuro. En el ejemplo del recepcionista, si el sistema es capaz de "desconectar" con un participante y comenzar la interacción con otro, únicamente diciendo "espera un momento" y retomar posteriormente la interacción con el primero de nuevo. En el caso de nuestro sistema de diálogo, si bien, actualmente no es capaz de mantener una conversación multiparte, sí que es capaz de pausar la interacción con frases como "cállate un momento" para retomarla con posterioridad con frases como "ya puedes seguir hablando".

TeamTalk [Marge et al., 2009] basado en *Ravenclaw-Olympus* facilita la interacción entre un operador y un equipo robótico. El operador propone una tarea dando diferentes comandos a los diferentes robots para localizar objetos específicos dispuestos en distintas localizaciones (el juego de la caza del tesoro). Sin embargo, el sistema sólo ha sido implementado en un entorno virtual llamado *USARSim* (*Unified System for Automation and Robot Simulation*).

4.2.9. Evaluación: escenarios y métricas

El desarrollo de nuevas teorías en interacción humano-robot y sistemas de diálogos y su implementación en robots reales es tan importante como su apropiada evaluación. En general, se puede considerar dos tipos de métricas: objetivas y subjetivas.

Las objetivas usan diferentes parámetros que dan una idea del nivel de satisfacción en la ejecución de una determinada tarea relacionada con el uso de diálogos. Algunos parámetros pueden ser: tiempo invertido, número de turnos intercambiados, número de malentendidos, número de subdiálogos de aclaración, porcentaje de errores de reconocimiento, etc.

Los parámetros subjetivos tratan de dar una idea sobre el grado de participación y comprensión del usuario, grado de entretenimiento, si la interacción le parece fácil y natural, si es eficiente, coherente, etc. James F. Allen distingue tres niveles de "*engagement*"[10]: inmersión, diagnóstico y fracaso. En la inmersión, el usuario se siente muy participativo y confortable con la conversación, siendo la comunicación natural y entretenida, como consecuencia se concentra únicamente en la consecución de la tarea asociada a la interacción, no en la interacción en si misma. En el nivel de diagnóstico, el usuario tiene problemas en la comunicación, pero intenta de manera cooperativa volver a la inmersión, por lo que se centra más que en el objetivo en el proceso comunicativo propiamente dicho. Finalmente en el nivel de fracaso, el usuario renuncia a seguir con la comunicación, sintiéndose incapaz y/o reacio a proceder con naturalidad, por lo que deja la comunicación o intenta obtener diversión con una "escena tan surrealista" (intenta obtener diversión de los propios malentendidos entre el sistema y él mismo, con el fin de lograr resultados inesperados y graciosos).

Las métricas usadas deberían depender de la población usada en el estudio: chicos, ancianos, expertos, usuarios sin experiencia, etc. Además los experimentos pueden hacerse con usuarios aislados o en grupo. Los resultados obtenidos de una evaluación normalmente repercuten en el diseño y refinamiento del propio sistema de diálogo.

Es importante no olvidarse, de que los métodos usados para la evaluación de los sistemas, deben estar muy bien especificados, y estructurados. Además es necesario un buen diseño de los escenarios concretos, para realizar un análisis riguroso del sistema.

Cuidado, rehabilitación y otras terapias

Un escenario típico en la interacción humano-robot es el que tiene que ver con el ámbito asistencial, en este sentido se han invertido fuertes cantidades de dinero, ya que puede ser una de los campos donde la robótica logre "despegar" haciendo mejor y más fácil nuestra vida diaria.

Care-O-Bot-II [Graf et al., 2004] ofrece algunas habilidades para el cuidado y

[10] participación/compromiso/acoplamiento

rehabilitación de ancianos. Este robot no implementa un DMS propiamente dicho, pero permite que el usuario se dirija a él mediante comandos de voz, para llevar a cabo tareas como transportar objetos en un ambiente doméstico. Su evaluación se hizo mediante simples cuestionarios.

Nursebot ([Pineau, 2003]) es un robot asistencial diseñado para ayudar a personas ancianas en sus hogares. El principal objetivo es que el robot actúe como enfermero y ayude en las necesidades diarias que le surjan al anciano. Su evaluación, en general, mide la viabilidad de asistir a personas mayores con limitaciones cognitivas y físicas en su día a día. Para medir la evaluación del sistema, propusieron dos métricas: tiempo en satisfacer una necesidad y porcentaje de errores en el diálogo. El gestor de diálogo usado estaba basado en la técnica de *POMDP*'s y *MDP*'s.

El DMS que está basado en *POMDP* y *MDP* presentado en [Roy et al., 2000] usa diferentes métricas para evaluar como de bueno es el sistema de diálogo. El principal parámetro tenido en cuenta es la recompensa total asociada al proceso de aprendizaje, acumulada durante todo el proceso de evaluación. La recompensa depende de como de bien y apropiada ha sido cada acción comunicativa (no olvidar que esta aproximación a la gestión del diálogo es puramente estadística).

En [Carnegie & Kiesler, 2002] se hace una evaluación de como las respuestas del humano se ven influenciadas por el diseño del robot y del propio diálogo usando un robot enfermero. El análisis de la varianza basado en el parámetro *ANOVA*[11] fue usado para medir características subjetivas y objetivas como: inteligencia del robot, personalidad, agradabilidad, consciencia, extroversión, neuroticismo, grado de apertura a nuevas experiencias, distancia de interacción, cuantos usuarios rieron o sonrieron, etc.

En [Kidd & Breazeal, 2008] se compara un entrenador personal para la perdida de peso mediante software de ordenador, mediante una tabla de ejercicios en papel, y finalmente mediante un sistema robótico diseñado con ese mismo objetivo. Los participantes llevan un seguimiento de su consumo de calorías y los ejercicios realizados. Su evaluación se centra en el porcentaje de uso del sistema, confianza de los usuarios en el mismo, afinidad entre ambos, etc. Para ello, se realizaron entrevistas, tests, y análisis en vídeo de las interacciones realizadas. Los resultados muestran que los participantes interactúan el doble de tiempo utilizando el robot que con los otros dos métodos, además de desarrollar una relación más estrecha con el robot. Ambos son indicadores del éxito a largo plazo de la pérdida de peso y el mantenimiento, y demuestran la eficacia de los robots sociales a largo plazo en HRI.

A pesar de que los robots sociales, con capacidades de interacción, han sido utilizados en estudios para medir como de útiles pueden ser en tareas terapéuticas con niños autistas ([Wainer et al.,], [Billard et al., 2007] y otros muchos), normalmente estos trabajos carecen de complejos DMS. Probablemente, debido al hecho de que la

[11] `http://www.ub.edu/aplica_infor/spss/cap4-7.htm`

enfermedad del autismo causa importantes dificultades a la hora de mantener una interacción natural con un sistema automatizado.

Guías en museos,centros comerciales, y otros centros públicos

Los museos han sido también espacios usados para el desarrollo de robots guías. Los experimentos permiten medir tareas a largo plazo y características, como por ejemplo, si el robot tiene alguna influencia en que los visitantes vuelvan a repetir la visita, media de tiempo de interacción, valores proxémicos como orientación y distancia del usuario frente al robot, el rol de gestos como la mirada o señalar, etc.

Por ejemplo, el proyecto Rovint [Lucas Cuesta et al., 2008] presenta un robot autónomo con un DMS basado en formularios que guía al usuario en un museo científico. El robot es capaz de establecer diálogos simples con los visitantes.

En [Kanda et al., 2010] un robot conversacional (no móvil) fue probado en un centro comercial, aunque se encontraba parcialmente teleoperado (el reconocimiento de voz no es automático, es llevado a cabo por humanos). Se diseño para proporcionar información sobre productos, precio y localización. Se evaluó durante 25 días en un total de 2642 interacciones. De las cuales 235 usuarios respondieron a unos cuestionarios. Con los resultados obtenidos se mejoró el sistema para dotarlo de mayor autonomía, y por lo tanto de menor cantidad de teleoperación, además se comprobó empíricamente que el comportamiento a la hora de que la gente comprase se veía influenciada por el robot.

El robot recepcionista Valerie, descrito en [Gockley et al., 2005], fue evaluado durante un periodo de 9 meses. Durante este tiempo se midió si la gente que interactuó con el robot, repitió posteriormente la visita y cuanto tiempo llevó cada interacción individual. El DMS del robot está basado en reglas, dado que inspira en los sistemas *AINE*, *AIML* y *ALIZE*([Wallace, 2000]), y usa más entrada de información por teclado que por voz. El dialogo llevado a cabo fue completar una biografía o diario de la estancia, por lo cual, las interacciones se repitieron durante este periodo de tiempo, pero fueron apenas de unos 30 segundos cada una de ellas.

El agente virtual descrito anteriormente ([Bohus & Horvitz, 2009]), que implementa un DMS multiparte, incluye algunas capacidades tales como detección y seguimiento de la cabeza y de la pose, gestión de la toma de turno, análisis de la escena e infiere algunos objetivos de los usuarios. Su evaluación mide como todos los componentes pueden funcionar conjuntamente, y el grado de atención de los usuarios en la conversación multiparte.

En [Looije et al., 2010] se evalúan tres asistentes para gente mayor con problemas de obesidad o diabetes: un robot real asistente (usando el robot *iCat* de *Philips*), un asistente virtual que simula el *iCat* y finalmente una interfaz textual. Comparando las ventajas e inconvenientes de cada uno de ellos. Su evaluación se hizo mediante

cuestionarios a los usuarios.

Robots que aprenden de, en, y para el entorno

En los trabajos de [Haasch et al., 2004] y [Toptsis et al., 2004]), un robot móvil se mueve por el entorno, obteniendo nueva información mediante las preguntas oportunas a los habitantes de la casa. En cambio, otros sistemas de diálogo, adquieren la información mediante etiquetación de corpus y relacionando palabras con objetos de una manera off-line, sin aprender en-vivo en la interacción con el usuario.

El diálogo en sistemas *IBL*, tables como el definido en [Bugmann et al.,] son evaluados comenzando con un análisis del corpus recogido, detectando las palabras clave para dar instrucciones y que permiten relacionarlas con el proceso de aprendizaje. Como *PAMINI* ([Peltason & Wrede, 2010b]) fue concebido para soportar distintas implementaciones y contextos de diálogo, de una manera similar a nuestro gestor de diálogo, en [Peltason & Wrede, 2010a] se hace una evaluación de como de fácil es implementar nuevos escenarios con diferentes características para cada uno de ellos, por lo tanto la evaluación está centrada en como el programador/experto desarrolla nuevas implementaciones de diálogo usando cada *DMS* y como de natural funciona con usuarios no entrenados. Un punto importante en *PAMINI DM* es que es independiente de la implementación del diálogo, característica fundamental de nuestro *DMS* aquí presentado (*IDiM*).

4.3. Contexto del diálogo

En esta sección se propone una nueva clasificación de los elementos que definen el contexto del diálogo que atiende a dos aspectos distintos: uno *extensional* y otro *temporal*. El primero se refiere a elementos que tienen una presencia estática, por ejemplo, las frases concretas del diálogo, el vocabulario manejado, atributos semánticos del reconocimiento, definición de los *actos comunicativos* que se utilizan, etc.

El aspecto temporal se refiere a elementos que influyen directamente en los cambios en el transcurso del diálogo: ritmos, silencios, intercambio de turnos, modulaciones no verbales, comienzo y finalización de un *acto comunicativo*, etc.

Esta clasificación, en estos dos aspectos de los elementos del contexto de diálogo, permite definir de manera sintética y sistemática lo que es un diálogo concreto en nuestro sistema.

4.3.1. Aspectos estructurales

El aspecto estructural del diálogo incluye tanto la definición del lenguaje formal que el robot es capaz de percibir, como la enumeración y definición de los AC's que es

capaz de expresar, así como la unión entre estos elementos del lenguaje con el modelo del mundo, y las capacidades, habilidades o acciones del robot.

Varios aspectos se ven involucrados, como por ejemplo, cómo el reconocimiento de voz es llevado a cabo. En el nivel fonético y morfológico la herramienta de reconocimiento de voz usa modelos acústicos y del lenguaje. En el nivel sintáctico se usa normalmente "gramáticas de contexto libre" (CFG), que establecen un conjunto de palabras y sus posibles combinaciones en oraciones válidas para ese contexto comunicativo; de esta manera se restringe notablemente las opciones comunicativas válidas para ese contexto. Todo ello será analizado en un capítulo específico, dedicado al reconocimiento de voz, en esta tesis.

Los niveles semánticos y pragmáticos[12], como se describe en [Morris, 1946],son más difíciles de formalizar e implementar en un robot social. Una de las soluciones adoptadas es incluir la parte semántica como reglas en la propia gramática (*CFG*) usada para el ASR ([Jurafsky & Martin, 2000]), pasando de ser una *CFG* a una *CFG semántica* (ver Fig. 4.1). Estas gramáticas establecen un conjunto finito de atributos y los diferentes modos posibles para asignar valores semánticos a ellos. Estos valores semánticos constituyen la información relevante para el sistema de interacción. Pero como se verá, cuando se hable de la *fusión multimodal*, no sólo el ASR es capaz de proporcionar valores semánticos al sistema.

También los gestos y otro tipo de acciones no verbales se pueden reunir en un conjunto discreto de elementos: un *gestionario*, que se describirá posteriormente. El gestionario incluye movimientos del cuerpo del robot como "afirmaciones", "saludos", "mostrar derrota", etc., pero también rasgos prosódicos como falta de fluidez, sonrisas, bostezos, alientos, etc.

Grosso modo, en los aspectos estructurales, se puede distinguir dos tipos de información: la información que es relevante para el robot obtenida mediante los *actos comunicativos* expresados por el usuario; y la información que el robot necesita expresar al usuario, mediante *actos comunicativos de expresión*.

Los *AC*'s que el robot es capaz de expresar son implementados de diversas formas, mediante "texto enlatado" para su síntesis mediante el sistemas de *TTS*, usando plantillas o "*snippets*" que realizan tareas básicas de generación de lenguaje natural (NLG), activando gestos, mediante expresión no-verbal de sonidos robóticos generados en tiempo real o pregenerados, o generando melodías de voz. Por ejemplo, el diálogo puede querer saludar, para ello puede activar el "*snippets*" que genera una frase de saludo que será sintetizada por el *TTS* con una emoción feliz; al mismo tiempo

[12]Las oraciones en sí mismas comportan un contenido semántico, pero su significado e interpretación adecuados no dependen sólo de ese contenido sino que requieren un contexto lingüístico definido para ser interpretadas (nivel pragmático). Es un hecho elemental bien conocido que una misma oración puede tener intenciones o interpretaciones diferentes en diferentes contextos (puede ser literal, irónica o metafórica).

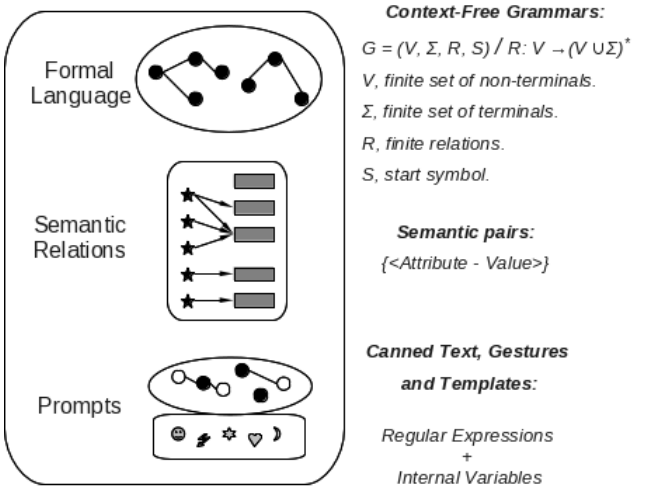


Figura 4.1: Aspectos estructurales de un proceso de interacción. Son los recursos necesarios para llevar a cabo la interacción, como son: las gramáticas semánticas, modelos del lenguaje, huellas de voz, gestionarios, diálogos, plantillas para la generación de texto no "enlatado", etc.

el diálogo activar el gesto de saludo (existe dentro del gestionario del robot); simultáneamente, y en un volumen menor, se generan sonidos no-verbales que también sugieran al usuario la sensación de ser saludado, mientras que en la tablet del robot se muestra alguna animación de bienvenida; de esta manera, para un mismo concepto semántico: "saludar", se ha "fisionado" en varias modalidades expresivas: voz, gestos, sonidos no-verbales y visuales.

4.3.2. Aspectos temporales

Como la comunicación natural es un proceso que sucede a lo largo del tiempo, se tiene que tener en cuenta sus características dinámicas, que constituyen los aspectos temporales del diálogo. El intercambio de turnos, la velocidad de oración, la duración y el significado de los silencios, son a los factores a los que nos estamos refiriendo. Además, existen otros factores dinámicos ha tener en cuenta por el diálogo, como son la distancia de interacción, el nivel de experiencia del usuario y los conocimientos que se aprenden dinámicamente.

Los aspectos temporales determinan, en gran medida, el proceso de *fusión multimodal*, es decir qué valores se empaquetan en un mismo *acto comunicativo*, cuáles se desechan y cuáles pertenecen al siguiente AC. Todo ello, se analizará detenidamente cuando se trate el módulo de *fusión multimodal*. A lo largo del proceso interactivo se comparten múltiples señales que van marcando en el tiempo el ordenamiento de

la información comunicativa. Existen así gestos no verbales que, por parte del que escucha, indican al hablante que está siendo atendido y comprendido, o no. Por ejemplo, el mirar a la cara o a los ojos (*gaze*), la afirmación con la cabeza, o sonidos suprasegmentales del tipo *aha*, *mhm*, silencios, etc.

La mayoría de los DMS siguen un modelo de interacción muy similar al de Shannon and Weaver [Shannon, 1948], donde hay un emisor que habla, y un receptor que escucha. Sin embargo, en una interacción natural es fácil encontrar solapamientos entre los interlocutores, además de que se produce un proceso de sincronización entre los interlocutores, en el que a la vez que hablan también escuchan, viendo y sintiendo lo que el otro usuario quiere expresar (interacción full-duplex).

El DMS intenta sincronizar los AC que el robot expresa con los los AC percibidos del usuario, siguiendo los siguientes parámetros:

- **Velocidad y Duración**: las oraciones y los gestos pueden ser más o menos rápidos.
- **Ritmo de Interacción**. Marcado por el parámetro anterior y por cómo se van sucediendo los *actos comunicativos* con los silencios.
- **Sincronía**: Medida de la adaptación de los ritmos de interacción de cada interlocutor.
- **Gestión del Silencio**: Cuando el silencio es mayor de un cierto umbral de tiempo, el interlocutor puede perder su turno. El silencio en sí mismo puede entenderse como un *acto comunicativo*, que exprese: duda, incoherencia, desatención, etc. [13]

[13] Una posible clasificación de las pausas: las pausas vacías o "silencios" y las pausas llenas o "titubeos". El silencio derivado de las pausas abarca un considerable espacio de tiempo del habla espontánea. Las interjecciones y los silencios corroboran la existencia de estas pausas. La investigación psicolingüística ha demostrado que las pausas vacías o "silencios" surgen durante las secuencias de desarrollo de las mismas cláusulas, es decir, se manifiestan en posición intraclausal. Por el contrario, las pausas llenas o "titubeos" constituyen pausas no silenciosas rellenadas mediante vocalizaciones como *eh* o *mmmm*. La mente humana aprovecha estos breves intervalos temporales como recursos estratégicos con el fin de diseñar mentalmente el discurso. Al mismo tiempo, algunas pausas esconden un propósito estilístico. Por lo tanto, puede advertirse que los fenómenos pausales desprenden múltiples funciones, fundamentalmente dos: la respiración y la planificación lingüística. La vacilación constituye también un fenómeno de planificación del discurso. En líneas generales, las pausas del habla espontánea reflejan las secuencias de desarrollo del proceso de codificación lingüística. Por consiguiente, puede afirmarse que las vacilaciones y las pausas representan disfluencias del habla espontánea. Cabe añadir que la investigación psicolingüística ha demostrado que una dosificación adecuada de pausas resulta incluso necesaria para el desarrollo de la producción del habla. *Fuente: Dept. Didáctica de la Lengua de la Universidad de Extremadura*

- **Intercambio de turnos**: Cómo se produce la petición o cesión del turno, así como el manejo de interrupciones del turno.

El intercambio de turnos ha sido y es objeto de estudio tanto en robótica, como en comunicación no verbal entre humanos [Birdwhistell, 1970], etc. En [Davis, 1971] se realiza una explicación minuciosa de cómo suele ser el intercambio de turnos. Cuando el que escucha quiere tomar el turno, comienza a moverse al mismo ritmo del que habla, el cual percibe la "solicitud". Ante esta petición, el que habla puede decidir ceder el turno o no. Si es que sí, entonces comienza a adaptarse él a los ritmos del que escucha dando así "el pie" y cediendo el turno.

La sincronía no solo se encuentra en el intercambio de turnos, sino que es parte constante del proceso de interacción presencial: ambos intervinientes sincronizan las velocidades e incluso las amplitudes de sus *actos comunicativos* ([Birdwhistell, 1970]).

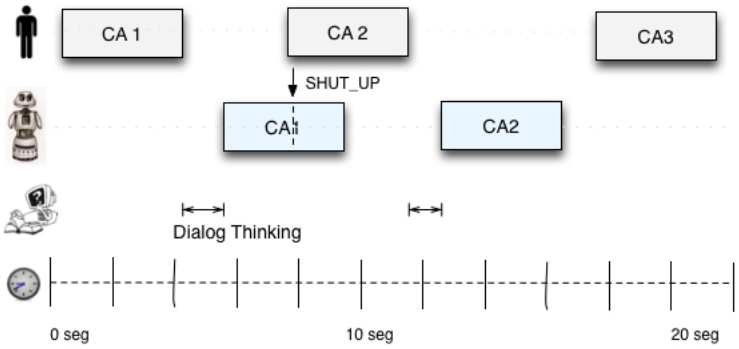


Figura 4.2: Aspecto temporal de un proceso de interacción.

En la Figura 4.2 se ilustra un escenario de intercambio de turno con interrupción (barge-in). El robot articula el CA_1 (acto comunicativo primero) pero en mitad del turno el usuario pierde el interés por lo que expresa el robot y comienza a articular su propio CA_2. Como el robot lo detecta, interrumpe su propio proceso de comunicación y transfiere el turno al usuario. Después de que el usuario finaliza su CA_2, el robot comienza a procesarlo (representado en la tercera línea como "*dialog thinking*"). El resultado de dicho procesamiento da como resultado un nuevo *acto comunicativo* por parte del robot. Pese a que el sistema de reconocimiento de voz es capaz de generar resultados parciales de reconocimiento [14] normalmente respeta el turno del usuario.

De manera natural, se ha observado que los seres humanos manejan un rango de tiempos de espera situado entre uno y siete segundos. Pasado este tiempo en el que

[14]El ASR puede entregar al DMS valores semánticos antes incluso de que el usuario haya dejado de hablar (anticipa los resultados tan pronto como los interpreta)

el hablante ha cedido el turno planteando una cuestión explícita o implícitamente, de manera natural el turno vuelve al hablante. También es habitual, que para no perder el turno, el que debe contestar suele emitir algún tipo de señal para mostrar que quiere mantener el turno, mientras elabora el *acto comunicativo* adecuado a la cuestión planteada. Por ejemplo, mediante señales suprasegmentales como "*mmmmm*", mientras elabora la respuesta.

4.4. IDiM: el gestor en RDS de diálogos interpretados

Nuestro DMS es implementado dentro de una arquitectura de control: AD-ROS. La gestión del diálogo se hace siguiendo el paradigma de rellenado de formularios, usando atributos semánticos y huecos de información ("*Information Slots*" (IS)). La Multimodalidad de entrada y salida está directamente unida al DMS. Esta sección termina con la descripción de cómo el sistema implementa nuestro modelo de diálogo basado en aspectos estructurales y temporales descritos anteriormente.

La relación que se establece entre los interlocutores marca el contenido de la interacción ([Paul Watzlawick, 1967]), por eso es importante establecer a priori cómo se espera que vaya a ser esta relación. En nuestra aplicación, los ejemplos de uso de *IDiM* que se van a presentar se basan en una relación entre el usuario y el robot que es complementaria: es el usuario quien, en principio, manda y ordena tanto al robot como al propio proceso de interacción.

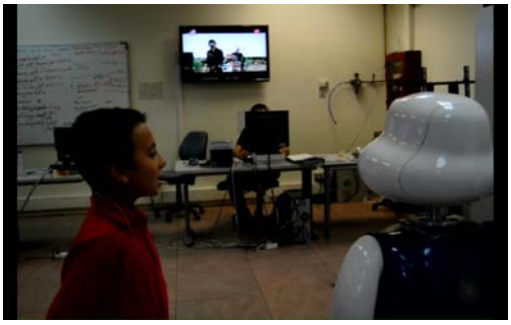
4.4.1. IDiM y la arquitectura de control del robot

La arquitectura de control del robot ha sido presentada en diversos trabajos. Su definición teórica la realiza Ramón Barber[Barber & Salichs, 2001b] y se muestra su aplicación en navegación. En el trabajo [Salichs et al., 2006b], se describe al robot Maggie que implementa esta arquitectura de control. AD es una arquitectura híbrida, basada en lo que se ha denominado como habilidades. Cada habilidad representa un módulo independiente de ejecución. Existen habilidades básicas o *primarias* que conectan directamente con los dispositivos de entrada y salida del robot: habilidades de habla (reconocimiento y síntesis), habilidades perceptivas como la que maneja el telémetro láser, habilidades de movimiento de las partes del robot y del robot en sí por el entorno, habilidad de tacto que hace de interfaz con sensores capacitivos en la carcasa del robot, habilidad de manejo del actuador de IR, etc. Dentro de este conjunto de habilidades básicas están las habilidades esenciales de comunicación verbal y no-verbal, que son utilizadas por el manejador de diálogo (ver Fig. 4.3).

Por otro lado, existen habilidades secundarias de más alto nivel que se sirven de las

(a) El juego del *Akinator* es una habilidad que interactúa con el usuario usando el tacto y la voz.

(b) Cuando la habilidad "Sígueme" está activada el robot te persigue.



(c) El usuario manejando la televisión a través del robot.

(d) El usuario puede leer algunas noticias de Internet.

Figura 4.3: El robot social *Maggie* en diferentes escenarios de interacción. IDiM permite cambiar dinámicamente el contexto del diálogo para cada habilidad activa.

habilidades básicas. Se tienen habilidades que son juegos, otras de domótica (manejo de una televisión por *IR*, o de luces mediante *RF*, etc), habilidades de movimiento (*followMe*, *teleoperation*, etc), juegos, habilidades de navegación, etc. La mayoría de estas habilidades, en algún momento, necesitan interactuar con el usuario. Dicho de otra forma, cada habilidad tiene sus propias necesidades comunicativas, y en consecuencia, cada habilidad ha de incluir su propia implementación de diálogo, usando el sistema de interacción aquí propuesto.

Es importante notar, la diferencia entre los diálogos, que pueden entenderse como la capa de interfaz o presentación, y las habilidades que podrían entenderse como la capa de "lógica de negocio". En la figura 4.4 se muestra el funcionamiento de IDiM. Una habilidad integrada en el sistema ("Secondary Skill" en la figura) que establece un conjunto de atributos semánticos que necesita para funcionar. En el caso de la habilidad que maneja la televisión, este atributo corresponde con el comando que se quiere enviar a la tele (véase figura 4.7). A su vez, la habilidad secundaria carga un diálogo en el DMS, donde se incluyen también no solo estos atributos semánticos, sino también acciones de Query y Filled.

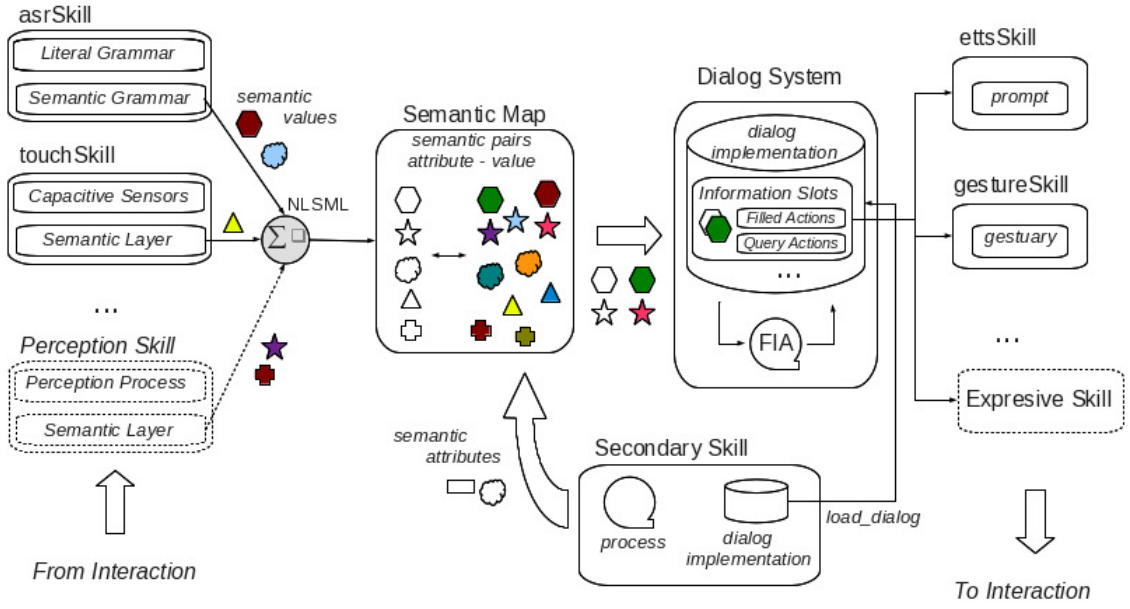


Figura 4.4: *IDiM* en un vistazo

4.4.2. Implementación y gestión del diálogo

IDiM distingue entre la implementación del diálogo, o simplemente *diálogo* y su interpretación o manejo. Por tanto, el sistema está formado por dos partes principales,

el manejador del diálogo y el diálogo en sí. Al separar el sistema general de diálogo (por sus siglas en inglés, DMS) en dos partes, una que interpreta y otra que implementa, se intenta flexibilizar los posibles diálogos que el sistema pueda mantener, ya que la ejecución del diálogo puede ir cambiando de contexto dinámicamente, según la evolución del flujo del diálogo.

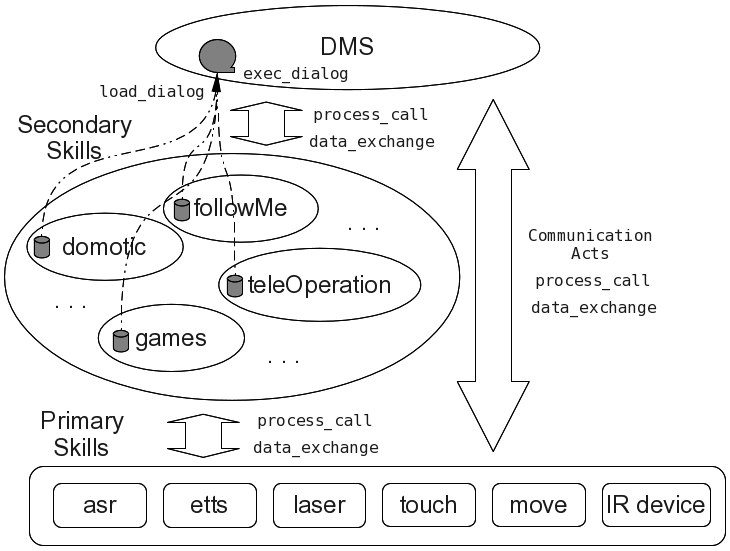


Figura 4.5: Cambios de contexto del gestor de diálogo. Cada habilidad necesita de diferentes diálogos. IDiM carga cada diálogo dependiendo de la habilidad que se encuentre activada.

En la figura Fig. 4.5 se representa la arquitectura del sistema. En la parte superior se muestra el DMS, que tiene el control central del robot. Cada habilidad secundaria que requiere interacción, implementa su propio diálogo. El DMS se encarga de cargar e interpretar en tiempo real distintas implementaciones del diálogo definidas por distintas habilidades.

En la figura se han representado las habilidades primarias más importantes: las habilidades de habla (*reconocimiento* y *síntesis* de voz), de movimiento, perceptivas (láser y tacto) y de control domótico de aparatos mediante dispositivos *wireless* (*IR devices*, por ejemplo). La ejecución, interpretación o manejo de los sucesivos diálogos que realiza DMS implica el envío de órdenes y compartir información tanto con las habilidades secundarias como con las habilidades primarias.

Por otro lado, se cuenta con un único manejador del diálogo que mantiene la misma política de diálogo para todas las habilidades que requieran de interacción verbal [15].

[15]mantiene el mismo algoritmo de rellenado de huecos de información para todos los diálogos

4.4.3. La fusión multimodal para *IDiM*

La percepción de la información para el diálogo es multimodal. Esto se realiza mediante *fusión multimodal* en una representación única de toda la información de cada *acto comunicativo* del usuario. Esta representación se implementa en el lenguaje estándar NLSML (Natural Language Semantic Markup Language)[16] y se verá en mayor profundidad en el siguiente capítulo.

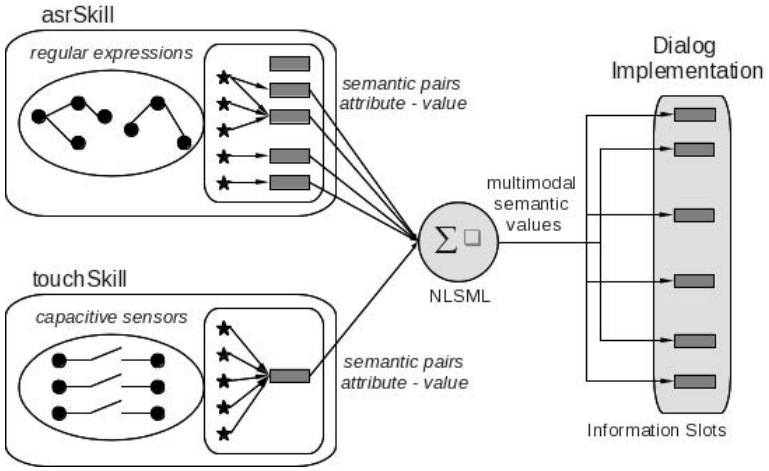


Figura 4.6: La fusión multimodal y el gestor de diálogo. Los pares semánticos atributo-valor son asignados directamente, uno a uno, a los huecos de información correspondientes.

En la figura 4.6 se presenta un esquema de *fusión multimodal* entre la habilidad *ASR* y la que capta los eventos de tacto de los sensores capacitivos del robot. Cada habilidad cuenta con un conjunto de atributos semánticos, que van tomando distintos valores según el funcionamiento de la habilidad. Todos estos pares atributo-valor se fusionan en una estructura *NLSML*.

Por ejemplo, si el usuario dice la expresión "*Maggie*, sube *este* brazo, por favor" mientras toca el brazo izquierdo, la *fusión multimodal* hará lo siguiente:

1. Primero se emite el evento `USER_SPEAKING`, que informa que el usuario está comenzado a hablar, y por lo tanto comenzando un *acto comunicativo*.

2. Mientras el usuario habla, el evento `TOUCHED` es lanzado por la habilidad de tacto, que informa sobre la parte del cuerpo del robot que ha sido tocada.

[16] http://www.w3.org/TR/nl-spec/

3. Cuando se acaba la locución y se reconoce un `REC_OK` será lanzado por la asrSkill, entregando valores semánticos de los atributos semánticos definidos en la propia gramática. El *acto comunicativo* finaliza.

Se creará una instancia de la información percibida cuando el *acto comunicativo* finalice incorporando como atributos semánticos los valores lanzados por las diferentes habilidades primarias, en este ejemplo: por la habilidad del tacto y la del reconocimiento de voz, pero otras muchas habilidades pueden estar involucradas, como son: detección de emociones, detección de usuarios, localización de usuarios, etc.

4.4.4. Atributos semánticos y huecos de información

La unión del sistema perceptivo multimodal con el diálogo se realiza relacionando uno-a-uno cada atributo semántico con un *Information Slot* dentro del diálogo. Esto afecta tanto al diseño de cada gramática semántica, como al resto de habilidades perceptivas, que en conjunto deben incorporar, al menos, todos los *Information Slots* de cada implementación del diálogo.

Sirva como ejemplo sencillo cómo funciona la habilidad que maneja la televisión mediante interacción con el usuario, tendría un *information slot* explícito: `command`, que puede tomar los siguientes valores:

- ON, para encender la televisión.
- OFF, para apagar la televisión.
- UP, para subir un canal.
- DOWN, para bajar un canal.
- FM, para cambiar a modo radio.
- DTV, para cambiar a modo televisión digital.
- V+, para subir el volumen.
- V-, para bajar el volumen.
- END, para dejar de usar la televisión

Este *information slot* solamente va a ser completado mediante la habilidad de reconocimiento de voz. Por tanto, se diseña una gramática CFG semántica relacionando los valores semánticos con el lenguaje formal que el robot es capaz de percibir (figura 4.7).

Así por ejemplo las siguientes frases distintas, provocan la misma asignación semántica, *commando* = *on*:

```
language es-ES;
tag-format <loq-semantics/1.0>;
public $root = [$GARBAGE] $telemando;
$telemando = ("enciende la television":On | "enciende la tele":On |
"pon la tele":On | "enciende":ON | "enciendela":ON |
"apaga":Off | "apaga la tele":Off | "apaga la television":Off |
"cambia de canal":up | "cambia el canal":up |
"cambia el canal de la tele":up | "sube un canal":up |
"cambia de canal":up | "baja un canal":down |
"pon la radio":FM | "quiero oir la radio":FM |
"pon la tele":DTV | "pon la television digital terrestre":DTV |
"sube el volumen":V+ | "baja el volumen":V- | "un poco mas bajo":V- |
"un poco mas alto":V+ |
"dejalo ya":ENDTV | "deja de manejar la television":ENDTV |
"deja la tele":ENDTV | "deja de manejar la tele":ENDTV){<@comando $value>};
```

Figura 4.7: Gramática para controlar la televisión mediante voz

- *"Maggie, por favor, enciende la televisión ."*
- *"Enciende la tele"*
- *"Maggie, pon la tele."*

En este ejemplo tan trivial puede verse la relación que existe entre el hueco de información que necesita la habilidad, el atributo semántico asociado y cómo se relacionan los valores semánticos con un lenguaje formal de contexto libre en el reconocedor de habla.

4.4.5. El gestor de diálogo basado en formularios y huecos de información

El sistema de diálogo descrito sigue el paradigma conocido como *"frame-based"*[17]. El diálogo se implementa en un lenguaje aumentado a partir de las especificaciones de `VoiceXML-2.1`[18]. Entre estas ampliaciones está el manejo de información multimodal, tanto en percepción como en expresión.

El control de la ejecución del diálogo está basado en el Form Interpretation Algorithm (FIA), que va recorriendo cada hueco de información y comprobando su estado. Los *actos comunicativos* que ordena el sistema de diálogo están modulados por distintos parámetros temporales definidos para cada diálogo concreto: tiempos de espera, velocidades de expresión, etc.

Tal y como se puede ver en la figura 4.8(a) la implementación de un diálogo, o "diálogo" a secas, consta de un marco (frame) que contiene los siguientes elementos:

[17] basado en formularios con huecos de información

[18] http://www.w3.org/TR/voicexml21

- **Huecos de información**: tienen que ser rellenados a través del diálogo. Pueden tomar dos estados: rellenandos o por rellenar.
- **Acciones de consulta o "Query Actions"**: son llevadas a cabo por el gestor de diálogo cuando un hueco de información no está relleno.
- **Acciones de completado o "Filled Actions"**: una vez que el hueco de información ha sido completado se ejecutan estas acciones.

El objetivo del DMS es completar todos los *Information Slots* presentes en la definición de un diálogo. Cada *Information Slot* se relaciona con un atributo semántico. El *FIA* (Form Interpretation Algorithm) se encarga de ir recorriendo cada Information Slot e ir comprobando si está o no vacío. Si el *IS* está vacío, el *FIA* ejecuta la *Query Action*. Está acción está orientada a provocar en el usuario un *acto comunicativo* que ayude a completar el hueco de información.

El sistema de percepción multimodal construye una estructura, donde se incluyen los atributos semánticos con sus valores, recibidos de la percepción del robot. Cuáles son estos atributos y el origen de sus valores es totalmente trasparente al *DMS*, de tal modo que un valor semántico puede venir tanto de información verbal como de la percepción de algún gesto del usuario, e incluso de percepciones internas de alguna propiedad del robot (por ejemplo el estado de carga de las baterías, la temperatura interna del robot, consumo de CPU, etc).

Las acciones asociadas a los dos estados del *IS* (vacío o completo) pueden ser de cuatro tipos. En la figura 4.8(a) se han clasificado según se refieran a acciones hacia el exterior o hacia el interior del sistema.

- **Acciones comunicativas**. En general cualquier *acto comunicativo*: frases, confirmaciones, gestos, aclaraciones, etc.
- **Acciones Ejecutivas**. Cualquier acción dentro de un dominio de acciones posibles que no son explícitamente comunicativas: obedecer la orden de un usuario, encender la TV, etc.
- **Movimientos en el flujo del diálogo**. En cualquier parte del proceso de interpretación de un diálogo, el robot puede cambiar el contexto del diálogo cargando una nueva implementación de un diálogo
- **Actualizaciones internas**. Rellenando variables globales como el nombre del usuario, o autorellenando otros huecos de información, así como activar otras habilidades.

Las acciones comunicativas y ejecutivas son externas porque actúan directamente en el entorno del robot. Las acciones internas se refieren a movimientos de un diálogo

a otro, o bien actualización de variables utilizadas por el propio diálogo o por las habilidades del robot.

En las asignaciones a variables internas dentro del diálogo también se considera poder realizar una asignación a un *IS*, dado que internamente también son variables del diálogo. Por ejemplo, imaginar los siguientes *IS* como siguen: `@body_action` y `@body_part` referidas a una acción de movimiento (subir,bajar, avanzar, mover, etc) y la parte del cuerpo involucrada (brazo derecho, izquierdo, cuello, etc). Ante el enunciado del usuario "levanta el brazo izquierdo" asrSkill completará ambos IS con los valores `@body_action = move @body_part = leftArm`. Sin embargo, ante el enunciado "avanza" puesto que la parte del cuerpo involucrada intrínseca es la base, automáticamente como acción interna dentro del diálogo se completará el IS `@body_part = base`.

La figura 4.8(b) muestra cada una de las partes de un diálogo para el ejemplo concreto de la habilidad que permite al usuario manejar la televisión mediante interacción con el robot.

En este ejemplo, se han escogido un conjunto de frases para la query-action, que provoquen que el usuario diga el comando que quiera enviar a la tele, o bien que informen de lo que puede decir. Frases del tipo *"dime un comando para la tele"*, *"si quieres apagar la tele dime apagar"*, etc.

Y como *filled-action*, el diálogo realiza el envío del evento oportuno a la habilidad primaria o de más bajo nivel, *IRskill*, que es la que maneja el dispositivo de emisión por infrarrojos. Se trata por tanto de una *Executive Action*.

4.5. Implementación de un diálogo

En esta sección se explican las principales etapas que se han de realizar para la implementación de un diálogo. Estas etapas son:

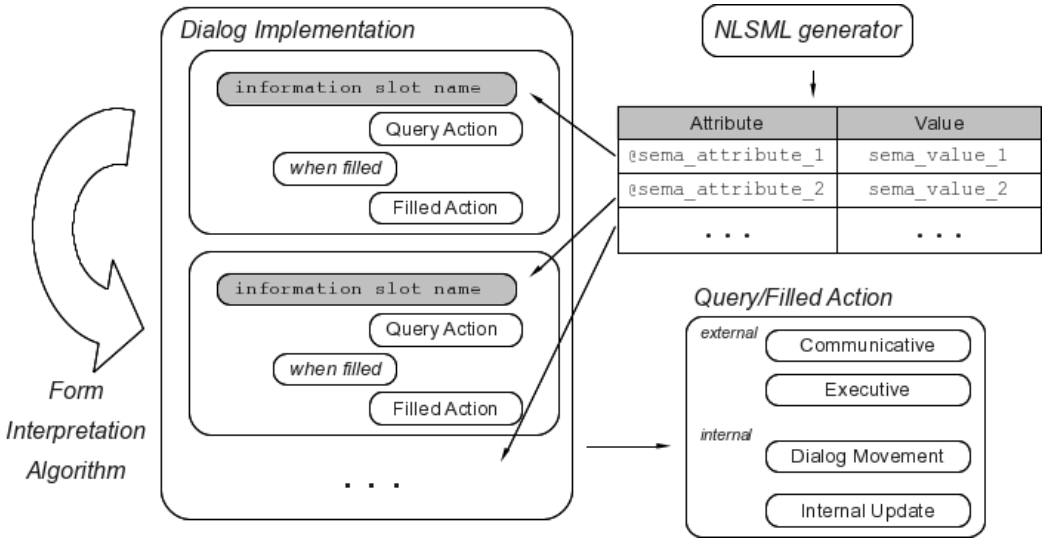
1. Diseño de las gramáticas para el reconocimiento de voz.

2. Diseño de los *actos comunicativos* de expresión o reutilización de los existentes, que permiten obtener la información necesaria.

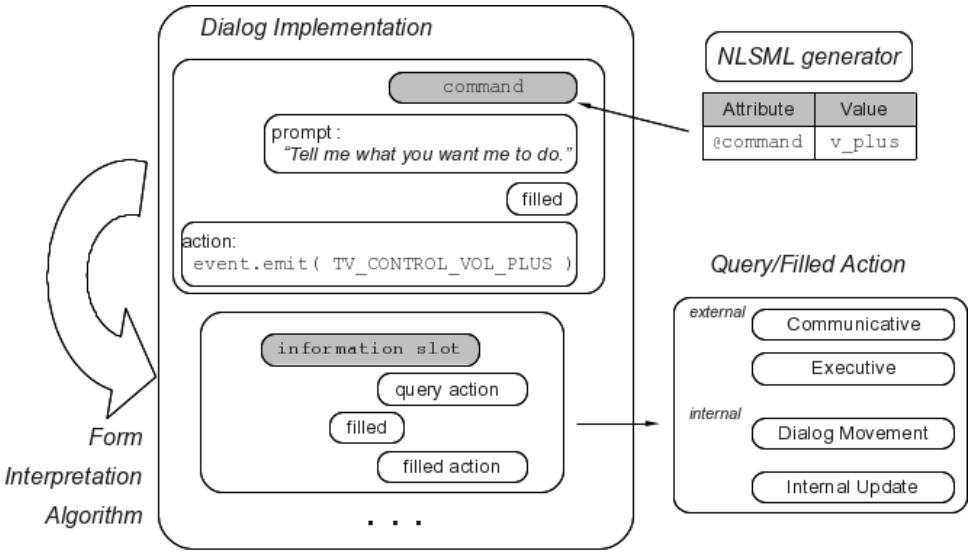
3. Incorporar tales *AC* en las "*Query Actions*" (acciones de consulta para conseguir rellenar los correspondientes huecos de información).

4. Identificación y enumerado de los atributos semánticos necesarios.

5. Incorporar tales atributos en las habilidades de percepción y en la implementación del diálogo.



(a) Esquema de implementación del diálogo. El diálogo realiza una Acción de Consulta ("Query Action"). Cuando un valor semántico completa un hueco de información, el diálogo realiza una Acción de Completado ("Filled Action").



(b) Interpretación y gestión de la implementación de un determinado diálogo.

Figura 4.8: Implementación del diálogo de control de la televisión

6. Diseño e incorporación de "*Filled Actions*" (acciones que se ejecutan cuando se rellena un hueco de información) en la implementación del diálogo.

7. Ajuste de los parámetros temporales del diálogo en su implementación.

De este modo se puede automatizar para todas las habilidades que requieran interacción, la creación del diálogo necesario. El potencial de esta automatización es tal que nos permite plantear la posibilidad de que sea el propio robot quien genere la implementación del diálogo, a partir de unas "necesidades semánticas" para una tarea, lo cual se propone como trabajo para un futuro a corto plazo.

4.5.1. Diseño del lenguaje formal

El lenguaje que el robot es capaz de reconocer se define mediante la construcción de gramáticas de contexto libre (*CFG*), a partir del conjunto de enunciados posibles por parte del usuario. Para la creación de estas gramáticas lo más habitual es proceder a la elaboración de un corpus de diálogos y de ahí obtener un modelo representativo. Por ejemplo, en [Bugmann et al.,] se hace un análisis exhaustivo de cómo las personas dialogan en un contexto específico para indicar direcciones entre usuarios.

La construcción de este corpus permite enumerar un lemario con las palabras que utilizan los usuarios, y se identifican las reglas gramaticales que las relacionan. También permite establecer características concretas de la representación del discurso, como por ejemplo, en qué puntos se hacen referencias a estructuras de la interacción ya enunciadas, etc.

La CFG permite incorporar los vocablos utilizados y las reglas gramaticales que los relacionan mediante una expresión regular ([Jurafsky & Martin, 2000] y otros). En las reglas de las gramáticas implementadas además se suele incorporar una variable, `$GARBAGE`, que sirve para representar vocablos del usuario fuera del lenguaje formal del robot. Para nuestro sistema, y diálogos implementados, se han añadido un conjunto de recursos genéricos y específicos (gramáticas, diálogos de ejemplo, etc).

4.5.2. Actos comunicativos de expresión

Los *actos comunicativos*, o también conocidos como *acciones comunicativas de expresión*, son parte de la estrategia de actuación comunicativa para la generación de un discurso coherente. Y, en cualquier caso, tienen como objetivo inmediato afectar a la conducta del receptor.

Dentro de nuestro sistema de diálogo, los *actos comunicativos* se utilizan dentro de las *Query Actions* y de las *Filled Actions*. Por ejemplo, a pesar de los datos que nos da la elaboración de un corpus, las faltas de entendimiento (*misunderstandings*) son parte habitual de la interacción natural. El robot, por tanto, debe ser capaz de actuar

adecuadamente para mantener la coherencia del diálogo y recuperar el entendimiento ante enunciados del usuario no reconocidos.

Por otro lado, los *actos comunicativos* sirven para hacer al robot más natural en su expresividad. Por ejemplo, se utilizan movimientos reactivos naturales como son bostezos, risas, suspiros, cosquillas, y demás *actos comunicativos* innatos en el ser humano.

4.5.3. Mapa semántico del dominio del diálogo

Cada habilidad interactiva establece un conjunto de atributos semánticos que son necesarios para la ejecución de la habilidad, y que conforman todo un mapa que se denomina "mapa semántico".

Por ejemplo, la habilidad domótica necesitará saber qué aparato va a manejar y qué comando necesita, luego como mínimo requiere de dos atributos semánticos. La habilidad que interactúa para conocer algunas características personales del usuario necesitará atributos semánticos como su nombre, su edad, su dirección de correo, etc. La habilidad que lee las noticias conectándose a un servidor de noticias necesitará saber qué tipo de noticias quiere el usuario: científica, deportivas, generales, información meteorológica, y así el resto de habilidades del sistema.

4.5.4. Atributos semánticos, acciones de consulta y acciones de completado

Cada atributo semántico requiere de uno o varios *Query Actions*, de distinto nivel de complejidad. Así, para obtener simplemente el nombre del usuario el robot puede preguntarlo directamente, pero para habilidades más complejas, el robot puede dar una breve explicación de qué es lo que espera percibir y de cómo lo tiene que explicar el usuario.

Un *acto comunicativo* del usuario (enunciado verbal o multimodal) puede completar varios atributos semánticos a la vez. El carácter semántico del atributo reside en cómo afecta su valor al robot, es decir, en su *Filled Action*. Así por ejemplo, el nombre del usuario se utilizará siempre que el robot quiera dirigirse al usuario de un modo más cercano, es decir, que su uso no depende de su valor. Sin embargo, el parámetro que define el tipo de noticias que el usuario quiere obtener influirá en cómo el robot realiza la búsqueda de dichas noticias.

Un "Filled Action" puede implicar autocompletar otros atributos semánticos que, por tanto, también ejecutarán su respectivas *Filled Actions*. Por ejemplo, la habilidad mediante la cual el usuario programa al robot una secuencia de acciones (véase [Gorostiza & Salichs, 2011]) si el usuario pide al robot que avance, se completará no

solo el atributo "@*action* = *move*", sino también, la parte del cuerpo involucrada "@*body* = *base*".

Un "Filled Action", también puede implicar un movimiento en el diálogo. Por ejemplo, cuando el usuario se registra o identifica, el robot cambia de contexto hablando sobre sí mismo e informando al usuario sobre sus posibilidades.

4.5.5. Parámetros temporales para el control del diálogo

IDiM es de iniciativa mixta, luego incorpora la propiedad *barge-in*, esto es, que el usuario pueda interrumpir al robot en medio de una locución. Antes de hablar el robot establece si la frase que va a decir es o no interrumpible.

Otra propiedad importante es el tiempo de espera de una locución del usuario. Cuando se supera este tiempo de espera el *Dialog Manager* emite un evento especial, y es la implementación del diálogo quien decide qué hacer ante este evento. Por ejemplo, en el "menú principal" si el usuario no dice nada, el robot cambia de estado a dormido.

4.6. IDiM respecto a otros gestores del diálogo

En la sección 4.2 se ha descrito los principales DMS aplicados a robótica hasta la fecha actual. Esta descripción ha sido estructurada dependiendo de varios factores. En esta sección se hace una breve reflexión sobre esos sistemas y cuál es la contribución que se hace en esta tesis con el gestor de diálogo IDiM.

La mayoría de los *DMS* desarrollados para robots sociales son específicos para un determinado contexto, es decir, se focalizan en obtener mediante interacción, algún tipo de información que esta fijada y no se puede cambiar en tiempo de ejecución. Por lo tanto, esos sistemas están fuertemente acoplados a un determinado contexto en un restrictivo dominio semántico. En ese sentido los *DMS* descritos anteriormente se enfocan en un asunto concreto de investigación: diálogos para el aprendizaje, diálogos multiparte, diálogos para actuar como enfermeros, etc. Se podría afirmar que hay una carencia general de independencia entre el *DMS* y la "implementación del diálogo".

También hay una carencia general de escalabilidad: la mayoría de los *DMS* presentados, especialmente los estadísticos (*MPD*'s, *POMDP*'s, etc.) tienen la enorme desventaja de que necesitan enormes corpus de diálogos etiquetados ([Gibbon, Moore, 1997]). Su entrenamiento se puede hacer mediante simulación, por ejemplo utilizando *HMM* (*Hidden Markot Models*), o usuarios reales, siguiendo un modelo de *Mago de Oz*[19] . Por todo ello, les hace difícilmente escalables a entornos

[19]Experimentos en los que el robot se teleopera por un operador, sin que el usuario tenga conocimiento de que el robot está siendo realmente teleoperado

comunicativos muy diversos, por lo que se suelen centrar en una tarea muy concreta (ej: obtener horarios de trenes...).

Resumiendo, esos DMS están altamente acoplados a una tarea específica, y no permiten cambios dinámicos en el contexto del diálogo, y en general no es nada sencillo el desarrollo de nuevos diálogos. Por contra, se ha tratado de proporcionar un nuevo sistema donde implementar un nuevo diálogo y unirlo con el resto de habilidades/capacidades de la arquitectura de control, es sumamente sencillo. Además todos los posibles diálogos gozan de un repertorio de información relativa a la interacción, a su alcance (encapsulada en cada *acto comunicativo*) que aporta información muy valiosa para la interacción y que se puede usar en cualquier diálogo sobre cualquier contexto comunicativo: localización de los usuarios respecto al robot, información sobre su edad, sexo, edad, colores de la ropa que lleva puesta, emoción con la que se expresa, pose, nivel de destreza con el sistema, parte del cuerpo del robot que está siendo tocada, etc.

Por otro lado, la mayoría de sistemas de diálogo funcionan como cliente o esclavo controlado por el sistema maestro que ejecuta un plan, siendo así el diálogo una parte más del plan. Así por ejemplo, cuando el robot necesita interacción con el usuario, el planificador activa el sistema de diálogo que sincroniza su información con la del sistema, y de manera subordinada interactúa con el usuario. El diálogo es así, una capacidad más del sistema que se activa y desactiva en ciertas situaciones.

Nuestro propósito es dar la vuelta a este modelo de control, intercambiando los papeles de maestro-esclavo. Ahora el sistema de ejecución de tareas es controlado por el sistema de diálogo. En este modelo, el diálogo puede actuar como control central (integrador de habilidades), teniendo la capacidad de ir activando o desactivando habilidades/módulos (mediante paso de mensajes) desde el propio diálogo, según avanza la conversación.

Nuestra arquitectura de control *AD-ROS* también concibe la hipótesis de un *sistema automático de decisión*, basado en un modelo de las emociones humanas que está fuera del objetivo de esta tesis, en el que este sistema de decisión, sí es el que toma el control, y se apoya en el Sistema de Interacción aquí presentado para interactuar con el usuario (ver [Malfaz & Salichs, 2004] y [Castro-González.,]). Incluso en el caso en el que el diálogo actúa como control central, activando/desactivando habilidades, se puede comunicar con las salidas del módulo de Toma de Decisión (la emoción del robot principalmente), para tenerlas en cuenta como una entrada más al sistema de interacción, de una manera similar a las entradas presentes en cada *acto comunicativo*.

No conviene olvidar la naturaleza multimodal del sistema [20], tanto en la entrada de información al sistema (después de la *fusión multimodal*), como a la salida del sistema

[20]El sistema presentado al completo corresponde al tipo de los llamados con simetría multimodal, esto es multimodalidad a la entrada y salida del Gestor de Diálogo.

(fisión multimodal). Pese a que nuestro sistema de diálogo se encuentra inspirado en el estándar VoiceXML, este estándar no está diseñado para una interacción multimodal; no obstante, tratando de mantener nuestros diálogos de una manera fiel al estándar, se ha incorporado de una manera elegante la manera de lograr la multimodalidad necesaria sin alterar gravemente el estándar, mediante el uso de un módulo de fusión [21].

El gestor de diálogo aquí presentado tiene varias fuentes de información que se describirán en las siguientes secciones del capítulo en detalle: la encapsulada en cada *acto comunicativo* (frase reconocida, pose, distancia de interacción...), la que obtiene del Sistema de Toma de Decisión (emoción y necesidades del robot), la que obtiene del "Perfil del Usuario" (edad, nombre, idioma, experiencia...), información directamente obtenida de servicios web (servicios de web semántica, traducción en-línea...). Todas estas fuentes de información posibilitan unos diálogos más naturales de los realizados por todos los sistemas vistos hasta ahora, ya que el programador del diálogo puede usar directamente en cualquier contexto comunicativo todas estas fuentes de información.

4.7. Resumen

Se ha presentado el gestor de diálogo *IDiM*, dentro del sistema *RDS*. La ventaja fundamental de *IDiM* reside en su versatilidad: en cómo el sistema permite cambiar de contextos de diálogo dinámicamente. Este sistema de diálogo se ha implementado para su uso en los robots social del grupo, especialmente el robot Maggie. Este robot cuenta con *un* conjunto muy amplio de habilidades, que además puede aumentarse en tiempo de ejecución. IDiM permite adaptar la capacidad de diálogo del robot a dicho conjunto de habilidades, manteniendo un único *DMS* para todas ellas con independencia de su número o función.

Se considera un avance haber formulado el diseño de un diálogo que tiene en cuenta tres aspectos o dimensiones: espacio-temporal, verbal y de relación; y que el diálogo pueda ser directamente implementado a partir de cuatro conjuntos de características: lenguaje formal, atributos semánticos, *query-actions* y *filled-actions*.

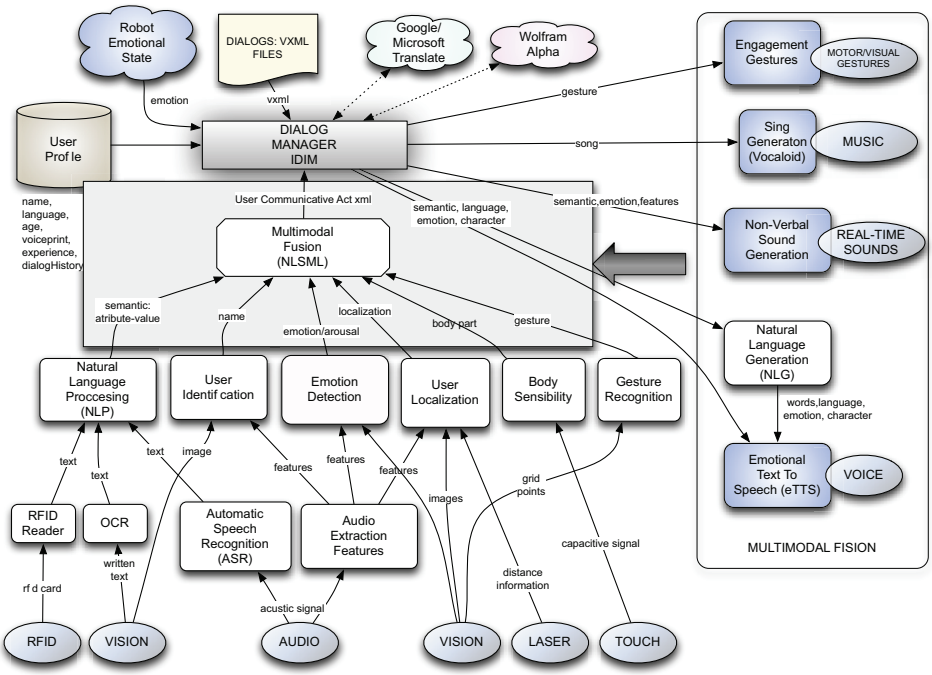
Naturalmente, el sistema cuenta con numerosas limitaciones. La primera y más clara es la falta de un modelo encargado de mantener la coherencia del diálogo en el cambio de dichos contextos. El robot pasa de una habilidad a otra, y en este paso la manera que tiene de conservar la información que ha ido adquiriendo no le permite realizar sus propias deducciones: falta un modelo de aprendizaje, que permita completar parámetros conceptuales del diálogo por inferencia, inducción o deducción.

[21] para profundizar más, ver la tesis sobre la que partió el trabajo mostrado en este capítulo [J.F.Gorostiza., 2010]

Sin embargo, el sistema es suficientemente robusto para ser utilizado por cualquier robot social cuya arquitectura esté basada en módulos de ejecución independientes que se han llamado habilidades, y en las que cada habilidad pueda tener asociada un contexto de diálogo concreto (parámetros conceptuales). Las pruebas experimentales con usuarios no expertos certifican la utilidad del sistema presentado. No obstante, para superar las limitaciones actuales se esta trabajando en el uso de un gestor de diálogo basado en planes.

CAPÍTULO 5

Sistema de fusión multimodal



"Todo debe simplificarse hasta donde sea posible, pero nada más."— Albert Einstein

5.1. Introducción: la teoría de actos comunicativos en la interacción entre humanos

Para entender el concepto de diálogo y multimodalidad, es necesario entender la teoría de *actos comunicativos* o *actos del habla*. Esta teoría se deriva del estudio de la comunicación verbal entre humanos [Bach & Harnish, 1979] [Bruner, 1975] [Searle, 1975, Searle, 1969], [Newcomb, 1953], en ellos se define claramente que los *actos comunicativos* son las unidades básicas del diálogo. Debido a estos estudios previamente citados, los ingenieros han encontrado que esta teoría es muy útil para describir y formalizar la comunicación que se puede realizar entre los humanos y las máquinas. Estos *actos comunicativos* permiten describir la interacción en interfaces de alto nivel de abstracción, de manera que en cada acto comunicativo se refleja una intención por parte del interlocutor.

5.2. Adaptación de la teoría de actos comunicativos al sistema RDS

Partiendo del estudio de las interacciones humanas y del análisis de los *actos comunicativos* como unidad de intercambio de información en el diálogo, se ha tratado de adaptar su uso al sistema de interacción aquí presentado.

Un usuario dialogando con un robot, lo hace mediante turnos que involucran comunicación verbal y no verbal. En cada uno de estos turnos se suministra cierta información que es relevante para el diálogo y que es necesario agrupar para que tenga significado temporal coherente. En cada turno, el usuario intenta transmitir un mensaje, que lo realiza mediante voz, adoptando una posición espacial respecto al interlocutor, una pose, con una determinada expresión facial, y usando un tono de voz de acuerdo a su emoción y al propio contenido del mensaje, etc. Toda esta información es extraída y manejada por el gestor de diálogo al agruparse temporalmente en un mismo mensaje comunicativo.

En la Figure 5.1 se representa este intercambio de información en un flujo de diálogo con varios turnos de conversación. El diálogo que se ilustra es un ejemplo concreto de diálogo producido con nuestro sistema, en el que tanto el usuario como el robot, en cada turno, tratan de transmitir un mensaje a su interlocutor. Este mensaje lo intercambian mediante los canales que consideren mas apropiados para sus propósitos. El usuario se comunica con el robot en el primer acto comunicativo (en la figura corresponde con la línea primera linea temporal), el sistema de diálogo tarda un cierto tiempo en procesar dicha información y preparar la respuesta, que normalmente es menor a un segundo (tercera linea temporal en la figura), entonces el robot responde con su primer acto comunicativo (segunda linea temporal); el usuario

recibe la información transmitida por el robot, pero antes de que finalice el acto comunicativo del robot, el usuario interrumpe al robot con un nuevo acto comunicativo, el robot interrumpe su mensaje y permanece callado hasta que el usuario finaliza su exposición. En este sentido, se habla de un sistema de diálogo *full-duplex*, en el que ambas partes pueden interrumpir el proceso comunicativo del otro interlocutor.

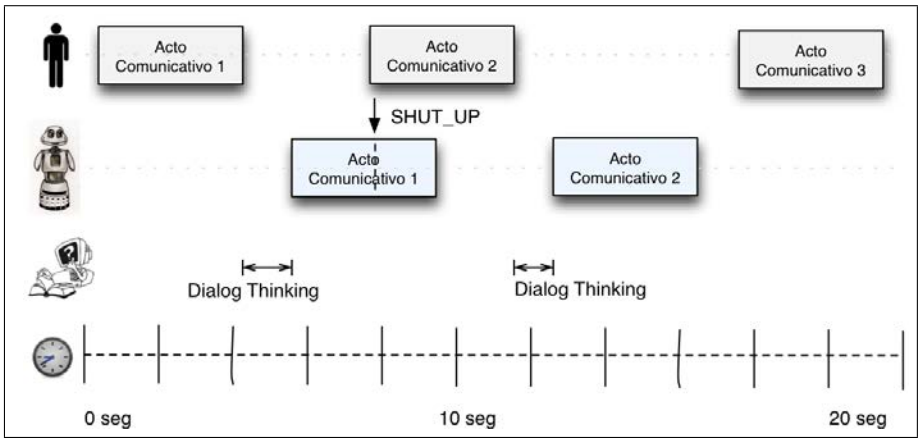


Figura 5.1: .
Ejemplo de diálogo full-duplex. Cada una de las partes puede iniciar la conversación y se pueden producir solapamientos e interrupciones en el turno de palabra

Para tratar de representar la información multimodal de manera coherente, el organismo W3C [1] ha propuesto un estándar que especifica mediante un fichero de texto XML como tiene que estructurarse dicha información. Este estándar se ha denominado NLSML (Natural Language Semantics Markup Language) [2]. Este estándar establece las reglas y la forma que tiene que tener el fichero XML donde queda reflejado las entradas multicanales recibidas, para que los gestores de diálogo basados en huecos de información puedan gestionarla.

El sistema de diálogo multimodal desarrollado permite interactuar con todos los robots sociales desarrollados en nuestro grupo de trabajo, aunque este trabajo se centre principalmente en el robot Maggie. En esta interacción, el usuario puede ordenar al robot que realice tareas para las que ha sido diseñado. El sistema de diálogo actúa como la capa de interfaz entre el usuario y el repertorio de "habilidades" del que está dotado el robot.

La interacción entre el usuario y el robot normalmente está guiada principalmente por el canal de voz, pero no es estrictamente necesario que sea así. La potencia del sistema multimodal permite al gestor de diálogo abstraerse del canal usado para la

[1] http://www.w3.org/
[2] http://www.w3.org/TR/nl-spec/

interacción, actuando de manera similar en el caso de que la entrada de información sea por voz, tarjetas de radio frecuencia o inclusive por texto escrito. De esta manera, se puede dialogar con el robot, en el sentido de intercambio de información, mediante tarjetas de radio frecuencia, con un pictograma ilustrado sobre las mismas, que indican la acción a realizar.

Hasta ahora, se ha comentado las ventajas de la multimodalidad relativa a la entrada de información por varios posibles canales. En este sentido, la complementariedad de la información recibida no se está teniendo en cuenta, ya que una de las posibilidades que nos brinda la fusión multimodal, gracias a esta complementariedad de la información recibida por distintos canales, es la de permitir resolver referencia deícticas (deixis), que son referencias a cosas que no están presentes en el contexto lingüístico. En este sentido, en expresiones deícticas como: "vete allí", apuntando con el dedo en una dirección, ese "allí" se resuelve mediante el componente que entrega al diálogo la pose con la que se expresa el usuario (la posición del cuerpo señalando a un punto espacial). En otro ejemplo de deixis, el usuario dice "levántalo", al mismo tiempo que toca uno de los brazos del robot, en este caso, la parte del cuerpo a resolver se resuelve mediante el módulo que entrega mediante el tacto la parte del cuerpo del robot tocada.

Dado que es necesario que la información recibida por los distintos componentes del sistema de diálogo RDS sean gestionados de manera coherente temporalmente. Por ello, se tiene que determinar en que instante empieza y finaliza cada acto comunicativo, teniendo en cuenta que cada módulo entrega la información semántica al módulo de fusión multimodal sin ningún tipo de sincronización. Es necesario analizar estos aspectos temporales para conseguir un correcto desempeño de un sistema de la complejidad que se acaba de presentar.

Figura 5.2: Etiquetas de radio frecuencia con pictogramas

5.3. Cuatro ejemplos de diálogos multimodales

En esta sección se va a describir lo principales escenarios donde se está usando y evaluando el sistema de diálogo para interaccionar de manera multimodal con usuarios no expertos con el mismo. En ellos, se ha podido analizar como realizar la fusión multimodal de la manera mas acertada posible. El primero de estos escenarios se corresponde con el diálogo principal del sistema que se encarga de gestionar la interacción entre el robot y el usuario, de tal manera que el primero de ellos actúa como esclavo del segundo, activando y desactivando habilidades según el usuario considere oportuno. En este escenario todas las entradas y salidas de información del sistema son usadas y fusionadas.

El segundo escenario permite controlar/teleoperar el movimiento del robot mediante voz. Está diseñado expresamente para probar la fusión multimodal (la salida del sistema ciertamente no es multimodal), y permite analizar cuando empieza y cuando finaliza cada acto comunicativo. En los dos siguientes escenarios, el reproductor de música y el diálogo en la nube, se presentan sistemas bastante novedosos, sencillos pero muy potentes, en los que fundamentalmente la fusión de información se realiza mediante el uso de dos reconocedores de voz, de propósitos diferentes, pero trabajando simultáneamente.

5.3.1. El integrador de habilidades por diálogo

El sistema de diálogo multimodal permite controlar al robot, de tal manera que este último actúa como esclavo del usuario satisfaciéndole, dentro de sus posibilidades de interacción. Para ello, el robot consta de un repertorio de habilidades o capacidades que es capaz de realizar, mientras que el diálogo es capaz de activar/desactivar dichas habilidades mediante interacción entre el usuario y el robot.

La arquitectura de control AD-ROS contempla las múltiples capacidades del robot en habilidades independientes. Cada habilidad puede tener dos estados principales: puede estar activa o bloqueada. Cada una de estas habilidades tiene asociada una implementación concreta del diálogo, y el conjunto total de habilidades se integra y se controla mediante un sistema de control central. Este sistema se ha implementado mediante un diálogo principal que se encarga de secuenciar las habilidades en tiempo real según vaya ocurriendo la interacción con el usuario.

Desde el diálogo principal (integrador), se puede transitar a otros subdiálogos (ver la Fig. 5.3). En este sentido, la especificación del diálogo puede verse como una máquina de estados o un grafo dirigido. Cada estado se corresponde con un determinado contextos comunicativo. Cada vez que se transita de un estado a otro, se produce un cambio en el contexto comunicativo. Estas transiciones consisten en cargar el nuevo subdiálogo correspondiente, con su/s gramática/s asociadas.

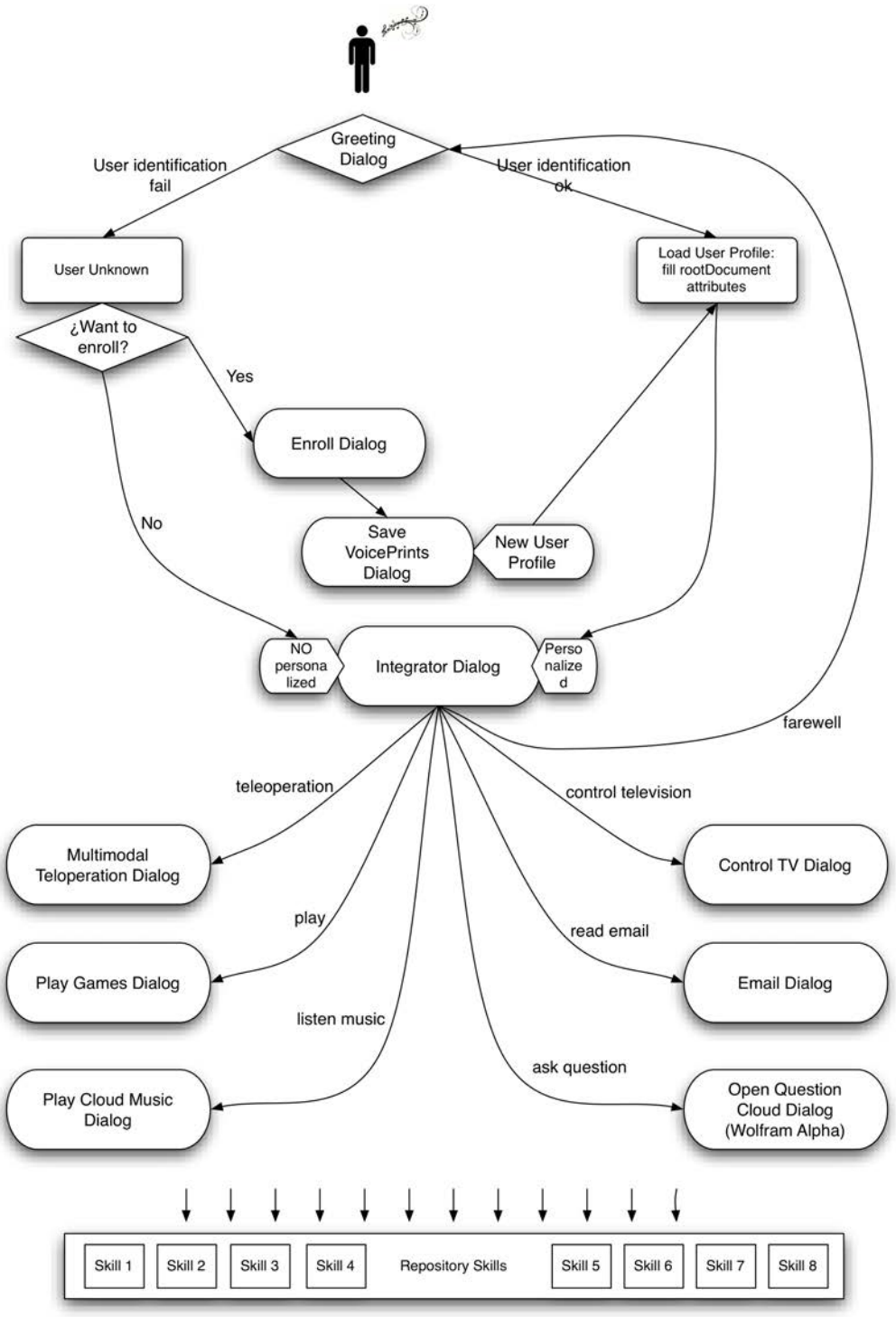


Figura 5.3: Ejemplo de diálogo como integrador de habilidades en el robot social Maggie. El robot actúa como esclavo, obedeciendo y ejecutando las acciones que ordena el usuario

.

La interacción comienza con el saludo del robot al usuario, mediante voz y sonidos no verbales. De esta manera el robot trata de atraer la atención del usuario, para iniciar una interacción. Si el usuario responde al robot devolviendo el saludo, se puede considerar que ha comenzado una interacción que es deseable mantener. Como se ha comentado en numerosas ocasiones la interacción es multimodal, de tal manera que el usuario puede saludar al robot no sólo mediante la voz, sino también mediante la etiqueta de radio frecuencia que representa el pictograma de saludar.

Cuando la interacción se realiza por voz, si el robot reconoce el timbre de voz del usuario automáticamente carga su perfil para adaptar mejor el diálogo a su interlocutor (idioma de voz usado, nivel de experiencia, estado emocional, preferencias, etc)[3]. Si por el contrario, el sistema no logra identificar al usuario, le pregunta si desea que se conozcan mejor. Si el usuario acepta, responderá a una serie de preguntas que le hace el robot, relacionadas con su nombre, edad, idioma preferido, etc. Durante este proceso de "registro" en el sistema, el robot aprende las características de la voz del usuario para futuras identificaciones o detección de emociones, así como sus preferencias personales. Si el usuario no desea "registrarse" en el sistema, la interacción continua pero sin ser adaptada al usuario.

La adaptación del diálogo al usuario se realiza en varios sentidos. Alguna de estas adaptaciones son las siguientes: el idioma en el que se conversa, la posición del robot respecto al usuario, más o menos diálogos explicativos del sistema en función del nivel de experiencia del usuario, la edad para sugerir o no ciertos juegos/habilidades, el nombre del usuario para dirigirse a él por su nombre además de poder consultar cuentas personales como la de su correo electrónico, el sistema de detección de emociones también se adapta al tono del usuario para aumentar la precisión de la clasificación, etc.

Una vez que el usuario ha sido identificado puede ordenar al robot que realice cualquier acción de entre las disponibles[4]. En este estado, el robot se encuentra en un modo de pregunta abierta, en el cual el robot actúa como un asistente capaz de obedecer a un conjunto bastante amplio de opciones disponibles. Actualmente el sistema tiene integrado unas veinte habilidades entre juegos interactivos, habilidades de información (dar las noticias deportivas o el tiempo meteorológico), reconocimiento de medicamentos mediante RFID, etc. No obstante, este dominio de habilidades es fácilmente ampliable e integrable en el sistema global.

El usuario puede ordenar al robot cualquier acción, por ejemplo su deseo de querer que le acompañe moviéndose por el entorno. Cuando se procede a activar una

[3]Se está trabajando en un sistema mas complejo de identificación de personas, en las que se combina la voz con aspectos visuales como: reconocimiento facial, de género e incluso de la ropa que lleva o llevaba puesta

[4]Sino ha sido identificado y no ha querido registrarse también llega a este estado del diálogo, pero sin adaptación

habilidad que pueda comprometer la seguridad del robot o del usuario, el robot pide confirmación. Por ejemplo, la habilidad *followMe* implica que el robot pueda moverse libremente por el entorno. Si el usuario confirma afirmativamente, el dialogo realiza la activación de la habilidad. A su vez, esta habilidad gestiona su propio subdiálogo que intercala síntesis de sonidos y de frases. El usuario puede desactivar la habilidad y volver al diálogo principal mediante otro comando de voz.

Volviendo al estado del diálogo principal que permite activar y desactivar habilidades del robot, el usuario puede, por ejemplo, pasar al subdiálogo que se encarga de controlar la televisión. Mediante este diálogo se puede controlar todas las funcionalidades que ofrece la televisión, y de forma general cualquier dispositivo controlarse mediante un mando infrarrojo universal. De la misma forma, el diálogo permite controlar cualquier habilidad de las que consta el repertorio de nuestra arquitectura de control.

Si el usuario desea finalizar la interacción, bastará con despedirse del robot o "mandarle a dormir". En ese momento, el robot se despide del usuario y transita al estado de "bienvenida" en el que puede establecer una interacción con cualquier otro usuario.

El diálogo desarrollado para la interacción con los usuario, hasta la fecha no es multiparte. Esto es, no es posible cargar más de un perfil de usuario, ni tampoco identificar varios usuario concurrentemente, ni separar la señal de audio recibida en varios canales de voz, por lo que no se puede mantener una interacción coherente y concurrente con varios usuarios simultáneamente.

5.3.2. El Diálogo de teleoperación multimodal

Centrándonos en la multimodalidad del sistema capaz de resolver referencias deícticas, se ha construido un diálogo que permite mover/teleoperar el robot para evaluar la fusión multimodal del sistema. Para ello, el usuario deberá indicar mediante cualquiera de los canales disponibles (voz,tacto, visión, RFID y gestos) la acción a realizar (Fig. 5.4). Veamos unos ejemplos de interacción gestionada por el sistema:

- El usuario dice "sube el brazo derecho". No toca ninguna parte del robot y no realiza tampoco ningún gesto con su cuerpo (pose neutra). Simplemente desea que el robot levente su brazo derecho. Para ello, toda la información necesaria está presente en la frase de voz: $Action : Up, BodyPart : Right_Arm$. Por lo tanto, mediante el uso de un único modo.

- El usuario dice "súbelo" al mismo tiempo que toca el brazo que desea que suba el robot. Es necesario que toque el brazo del robot antes de que se finalice el reconocimiento de la frase pronunciada, sino el sistema no sabría que parte del

cuerpo tiene que subir. En este caso nuevamente los huecos de información son: $Action : up, BodyPart : Right_arm$. Ver figura 5.4 a).

- El usuario dice "gíralo a la derecha" al mismo tiempo que le toca la cabeza. En este caso el usuario desea que el robot gire su cuello a la derecha: $Action : Turn_left, BodyPart : Head$. Ver figura 5.4 b).

- El usuario dice "gira la cabeza a la izquierda" sin tocar ninguna parte del robot. Este caso es equivalente al anterior, sin embargo toda la información necesaria está contenida en la frase de voz, esto es, no hay deíxis. $Action : Turn_left, BodyPart = Head$.

- El usuario dice "vete en esa dirección" indicando con su brazo la dirección a la que quiere que se mueva el robot. En este caso la pose del usuario indica la dirección del movimiento. $Action : Move_Left, BodyPart = Base$. Ver figura 5.4 c).

- El usuario dice "vete a la derecha" sin tocar ni indicar con ningún gesto nada. Toda la información está presente en al frase de voz: $Action : Move_right, BodyPart : Base$.

- El usuario dice "vete al cargador" sin tocar ni hacer ningún gesto. En este caso, el robot se irá a la posición donde se encuentra la estación de carga par recargar sus baterías. En este caso nuevamente toda la información necesaria se encuentra en la frase de voz, no obstante, para saber la posición del cargador se usan las habilidades de navegación geométrica presentes en el robot, que le permite moverse hacia sitios conocidos una vez se ha localizado.$Action : Move_chargeStation, BodyPart : Base$. Por lo tanto, uno de los huecos de información se rellena mediante el acceso al modelo del mundo del robot.

- El usuario dice "vete a la puerta" sin tocarle ni hacer ningún gesto. Nos encontramos en un caso similar al anterior. $Action : Move_door, BodyPart : base$.

El diálogo concreto encargado de realizar la teleoperación se especifica en un fichero XML que actúa como capa de interfaz entre los sensores (mediante fusión) y de los actuadores (mediante fisión). En dicho dialogo se tienen tres huecos de información a rellenar: acción a realizar, parte del cuerpo involucrada y dirección de dicha acción.

En la figura 5.5 se puede ver un diagrama con la estructura conceptual del dialogo aquí presentado. Obsérvese que no todas las entradas multimodales son tenidas en cuenta. Las salidas multimodales del sistema son los gestos motrices involucrados en la intención del acto comunicativo: desplazarse por el entorno, mover una extremidad

(a) Up the right arm (b) Turn the head right (c) Go there

Figura 5.4: El usuario da ordenes de movimiento al robot mediante voz, o una combinación de voz, gestos y/o tacto

arriba o abajo, girar la cabeza a izquierda o derecha y hablar para comunicar la acción a realizar.

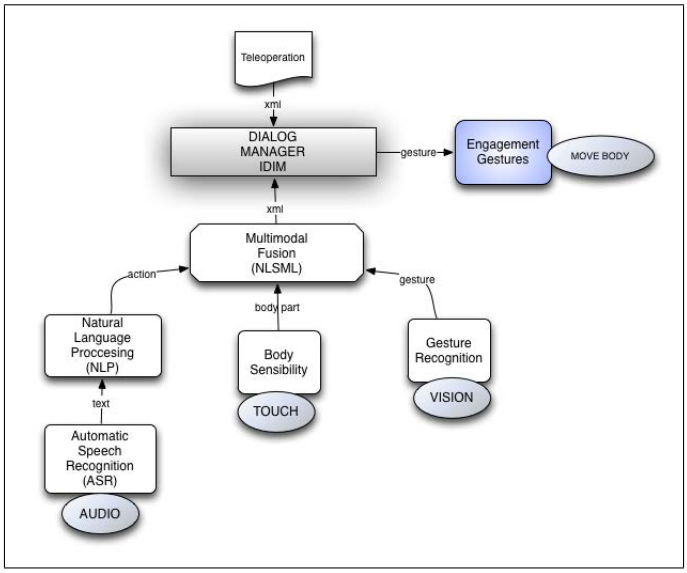


Figura 5.5: Esquema del diálogo de teleoperación

Para el relleno de estos huecos de información, el módulo que aquí se está describiendo de fusión multimodal debe saber como realizar dicha fusión en *actos comunicativos*. Para ello es necesario establecer cuando empieza y finaliza cada acto comunicativo. Sirviéndonos de nuestro pequeño diálogo para teleoperar el robot y analizando la evolución temporal de los eventos de resultados semánticos disponibles por cada uno de los módulos que entregan información al módulo de fusión multimodal, se ha determinado que cada acto comunicativo queda separado por la notificación

de final de reconocimiento de voz. Es decir, el evento que entrega el ASR (Automatic Speech Recogniton) cuando el reconocimiento de voz ha concluido servirá para marcar la frontera entre diversos *actos comunicativos*. En el caso de tener varios motores de reconocimiento ejecutando de manera concurrente (Loquendo y Google ASR), se espera por la obtención de los resultados de reconocimiento de ambos motores, antes de dar por concluido el acto comunicativo.

Si un mismo evento, con distintos valores semánticos es lanzado durante el mismo acto comunicativo, esto es, antes de la notificación de final de reconocimiento, será la información suministrada en el último evento de ese tipo, la que se tenga en cuenta. En figura 5.6 se puede ver una secuencia temporal de como se realiza la fusión multimodal en *actos comunicativos*. Los valores que reconoce cada uno de los módulo se entregan al módulo de fusión multimodal en cuanto están disponibles. Esto implica que la llegada de estos valores se hace de una manera asíncrona y desincronizada, siendo el componente de fusión el que los agrupa en un mismo acto comunicativo. Este acto comunicativo queda reflejado en un fichero de texto XML, que es entregado al gestor de diálogo IdiM, para que de manera coherente trate esta información fusionada temporalmente (en la figura, aparecen a derecha). En la figura se representan dos *actos comunicativos*, en el primero de ellos, el usuario toca el brazo derecho del robot y le dice que quiere que lo suba. En el segundo acto comunicativo el usuario toca la cabeza del robot y le dice que la gire a la izquierda. Son solo dos sencillos ejemplos para ilustrar como se realizar la fusión multimodal, pero cualquiera de los enumerados previamente son posibles.

5.3.3. Diálogo de reproducción de música en línea

Otro diálogo de ejemplo, que ilustra las posibilidades multimodales de nuestro sistema de diálogo y que resultaría en una utilidad muy interesante para un posible usuario, es el que permite que el robot pueda reproducir la mayoría de las canciones de cualquier grupo que se pueda encontrar en Internet. El usuario únicamente debería decirle al robot la canción, el grupo, o ambas cosas a escuchar, el robot casi instantáneamente comienza a reproducir la canción o las canciones solicitadas. En este caso, la fusión multimodal unifica la información de los dos motores de reconocimientos disponibles en nuestro sistema, el basado en gramáticas y el de gramática abierta (de texto libre), cuya finalidad y uso es diferente. Por otro lado, se logra incorporar un nuevo tipo de salida al sistema como es la reproducción de canciones.

Para su implementación, se ha usado el servicio de reproducción e intercambio

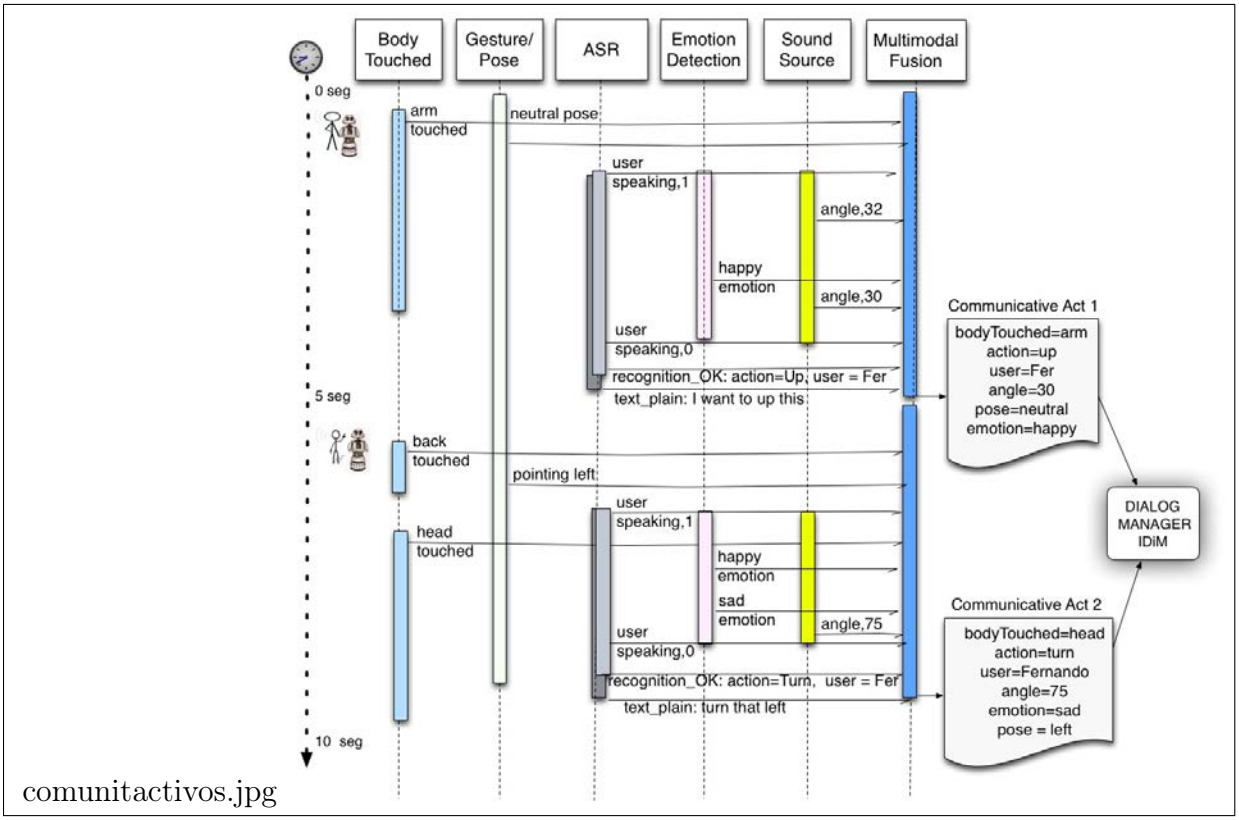


Figura 5.6: Ejemplo de la fusión multimodal de dos actos comunicativos

musical Goear [5] y un script[6] en lenguaje Perl [7] que sirve como interfaz para reproducir una canción o varias de un mismo grupo musical. Esto permite que el propio sistema de diálogo pueda acceder a los servicios del portal musical.

En este ejemplo se ilustra la fusión del reconocimiento de voz basado en gramáticas (mediante Loquendo ASR) y el reconocedor de texto abierto (mediante el servicio web de Google ASR). Cuando el usuario le dice al sistema que quiere escuchar música está rellenando un hueco de información, basado en gramáticas, que corresponde con la acción que se quiere realizar. Por otro lado, el diálogo se mueve a otro formulario, donde un hueco de información es "TextPlain" que representa la canción o el grupo a reproducir. Este texto plano es rellenando mediante el reconocedor de voz, no basado en gramáticas, de Google ASR. Este reconocedor de voz no genera resultados

[5] http://www.goear.com

[6] "Un script es un guión o conjunto de instrucciones. Permiten la automatización de tareas creando pequeñas utilidades. Es muy utilizado para la administración de sistemas UNIX. Son ejecutados por un intérprete de línea de comandos. Usualmente son archivos de texto.

[7] http://www.splitcc.net/Downloads/playgoear.pl

semánticos, sino que únicamente nos devuelve la transcripción textual de la frase pronunciada verbalmente. Si se quisiera rellenar el hueco de información, relativo a la canción o el grupo a escuchar, mediante un reconocedor basado en gramáticas (como el de Loquendo), se hubiera necesitado una gramática de gran tamaño que recogiera todas las posibles canciones y grupos a escuchar. En la práctica, constituye una solución casi inviable, por su alto coste y dificultad de mantenimiento.

5.3.4. Diálogo de pregunta abierta

En este escenario el robot es capaz de responder a casi cualquier pregunta que el usuario pueda formularle mediante voz y en cualquier idioma. Preguntas como por ejemplo:

- "Cual es la distancia de la tierra a la luna"
- "Donde estoy"
- "Cual es la capital de Francia"
- "Que sabes sobre Rafael Nadal"
- "Que tiempo hace hoy en Nueva York"
- "Cuanto es 12 x 3"

En este caso se ha diseñado un sistema de diálogo muy similar al famoso Siri [8], que es capaz de responder a preguntas complejas.

El usuario primeramente deberá decir al sistema que quiere realizarle una pregunta y a continuación la pregunta concreta a realizar; el robot le repite la pregunta realizada y a los pocos segundos contesta por voz con la respuesta más apropiada posible.

Internamente el sistema de diálogo funciona de manera similar al escenario anterior, primero rellena el hueco de información de acción a realizar, mediante Loquendo ASR y se mueve a un nuevo formulario donde se quiere rellenar la pregunta concreta a formular, mediante un hueco de información de texto plano. Este hueco de información se rellena mediante el reconocedor de Google ASR que es capaz de transcribir en multitud de idiomas la pregunta formulada.

Si se está dialogando con el robot en un idioma distinto del inglés, se usan los servicios de traducción en línea de Google o Microsoft para traducir dicha pregunta al idioma inglés. Dicha pregunta en inglés, es enviada a la base de conocimiento

[8] http://www.apple.com/iphone/features/siri.html

Wolfram Alpha [9] que responde a dicha pregunta, con la información que es capaz de obtener de su base de conocimiento, mediante un fichero XML.

Una vez se tiene la respuesta, se interpreta el fichero XML para obtener las partes más interesantes de dicha respuesta (principalmente aquellas que se pueden comunicar por voz), y son sintetizas por el sistema de diálogo en el idioma en el que se estaba dialogando, siendo nuevamente necesario una traducción online si el idioma era distinto del inglés. Si las traducciones no son correctas se puede perder información valiosa para el diálogo, también hay ciertas preguntas para las que el sistema no tiene ninguna respuesta, en esos casos se le comunica al usuario que no se dispone de dicha información, en diferente posibles expresiones (no siempre dice la misma respuesta).

Otro aspecto que se ha tenido en cuenta, para que el sistema sea lo suficientemente interactivo, es la velocidad de respuesta del diálogo. Para ello, se ha implementado a bajo nivel un sistema de detección de comienzo y fin de voz, para enviar al servicio de Google ASR única y exclusivamente las muestras de voz en cuanto estas están disponibles y en un formato de audio comprimido (flac o speex), de esta manera se obtiene la transcripción en apenas un segundo de retraso desde que se finalizó la locución. A su vez el servicio "Wolfram Alpha" y el parseo de la respuesta demora unos 3 o 4 segundos, tiempo que es compensado con la síntesis, por parte del robot, de la pregunta que ha formulado el usuario antes de comunicarle su respuesta.

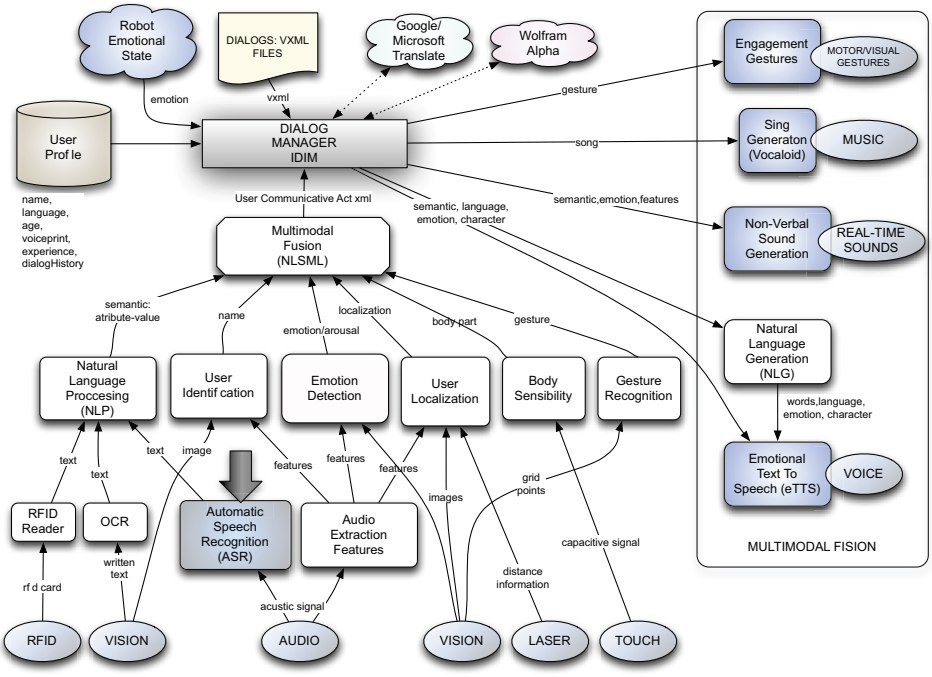
5.4. Resumen

En este capítulo se ha descrito como se realiza el proceso de fusionar el conjunto de entradas multimodales, en un único paquete de información coherente en el tiempo, capaz de ser interpretado por el gestor de diálogo. El trabajo desarrollado se ha inspirado en la teoría de *actos comunicativos* en la interacción entre humanos. Finalmente y mediante ejemplos, se ha descrito y analizado los principales diálogos implementados para este trabajo y que sirven como pruebas experimentales para mostrar esta fusión multimodal.

[9] http://www.wolframalpha.com/

CAPÍTULO 6

Sistema de reconocimiento automático del habla



"Asegúrate antes de hablar que tus palabras son mejores que el silencio."— viejo proverbio árabe

6.1. Introducción

Cuando se habla de interacción-humano robot, se asume que el robot tiene la capacidad de reconocer las palabras que el humano transmite verbalmente de manera precisa, incluso entender el significado de dicha oración, es decir, de extraer la semántica asociada a dicha frase.

Los humanos tenemos cinco sentidos básicos – visión, tacto, oído, olfato y gusto- con los que percibimos nuestro entorno y nos hacemos una concepción mental del mundo. Sin embargo, puede resultar paradójico, que siendo el habla el medio de comunicación por excelencia entre las personas, la interacción con los equipos informáticos se lleve a cabo casi siempre mediante teclado, ratón o pantalla táctil[Jansen & Belpaeme, 2006].

Las denominadas tecnologías del habla tienen como finalidad facilitar el uso de los ordenadores, en nuestro caso de los robots, introduciendo y recibiendo información de modo oral. Para alcanzar este objetivo es necesario contar, al menos, con tres tecnologías básicas: las que posibilitan que la información escrita se convierta en voz – síntesis del habla-, las que permiten que un sistema informático realice las tareas que se le solicitan verbalmente – el reconocimiento del habla- y las que facilitan la interacción oral entre una persona y un servicio – los sistemas de diálogo - que se basan en las dos tecnologías previamente citadas.

Por reconocimiento automático del habla (automatic speech recognition o ASR) entendemos el proceso de capturar y convertir una señal acústica, recogida por un micrófono, en texto escrito, mediante un computador. Las tecnologías de ASR se dividen en dos tipos básicos: uno es "*speaker-dependent system*", en el que el sistema está entrenado para un usuario en concreto y sólo es capaz de reconocer adecuadamente la voz de ese usuario (el reconocimiento es abierto, todas las frases son posibles y válidas; se les suele llamar herramientas de dictado). El otro tipo es "*speaker-independent system*" el cual es capaz de transcribir las oraciones enunciadas por cualquier usuario y sin entrenamiento previo con el sistema. En el campo de la HRI el interés se centra principalmente en el segundo tipo de sistemas.

Según Llisterri [Llisterri et al., 2003], Liddy [Liddy, 2005] y Feldman [Feldman, 1999], en cualquier sistema con reconocimiento automático del habla existen los siguientes niveles lingüísticos de abstracción: nivel fonético, morfológico, léxico, sintáctico, semántico, de discurso, y pragmático. Estos niveles abarcan la construcción de palabras como conjunción de fonemas; las reglas válidas de posicionamiento de las palabras dentro de una oración; la extracción del significado semántico de la oración en el contexto comunicativo y en relativas al discurso general. Este capítulo se ha centrado en los primeros niveles: léxico, sintáctico y semántico, dejando a un lado el nivel pragmático o del discurso, en los que actualmente se está trabajando y se está próximo a realizar una publicación al respecto.

En los estudios sobre HRI encontrados en la literatura, no es habitual que se facilite información sobre el software usado para realizar el reconocimiento automático del habla, ni tampoco los sistemas físicos de captación del audio. Se pueden encontrar algunos estudios, como por ejemplo [Ishi et al., 2006b][Kim & Choi, 2007][Kibria & Hellström, 2007], que se centran en robots móviles, por lo que el aspecto social pasa a un segundo plano. La mayoría suelen trabajar únicamente en idioma inglés, lo que hace muy difícil extraer y extender conclusiones de su uso con otros idiomas. Por último, no se hace un estudio completo, comparativo y preciso de la calidad del reconocimiento con las distintas opciones de software/hardware actualmente disponibles y en diversos entornos.

En esta sección se persigue el objetivo de dar las pautas y una visión general de los aspectos y tecnologías a tener en cuenta a la hora de dotar a un robot de la capacidad de entender el lenguaje natural hablado. Nuestro objetivo no ha sido desarrollar desde cero los algoritmos y los componentes básicos para tal función, sino la de integrar las tecnologías existentes con nuestra arquitectura de control para conseguir la más alta calidad posible en la HRI por voz.

Otros trabajos similares que se están desarrollando y de relevancia son: la librería open-source HARK [Nakadai et al., 2008a], que incluye los módulos de localización de fuente sonora, ASR y reconocimiento multicanal. Otros estudios recientes y que abren una nueva vía de investigación, son los que fusionan la información visual de lectura de labios, con la sonora, para mejorar el reconocimiento global [Yoshida et al., 2009].

Existen numerosos robots con la capacidad de realizar reconocimiento automático del habla. Se puede hacer una división en dos tipos de robot con esta capacidad: los primeros de ellos son los llamados “chatbot” o “robots virtuales”, los cuales carecen de cuerpo físico real, y su interacción es a través de un computador tradicional. Por otro lado, están los robots dotados de un cuerpo real, entre ellos se puede mencionar algunos de los más relevantes para la HRI, como son Honda ASIMO [Sakagami et al., 2002], SIG2, Robovie[Mitsunaga et al., 2006], HRP-2 [Takahashi et al., 2010], todos ellos dotados del software HARK. IROBAA [Kim & Choi, 2007] es capaz de localizar la fuente sonora mediante la fusión de la información visual y auditiva. JIJO-2 [Fry et al., 1998], es capaz de convivir con humanos en un ambiente doméstico, al igual que Robovie, siendo capaz de aprender el nombre de ciertos objetos y localizaciones por voz (short semantic memory). HERMES [Bischoff & Graefe, 2004] es un robot capaz de entender el lenguaje natural, el cual ha sido probado durante 6 meses como guía de museo. Todos estos robots, son capaces de entender el lenguaje natural hablado, pero sólo en un subconjunto del inglés y/o Japonés.

A continuación, en la sección 6.2 se describen los mecanismos hardware para captar el sonido y la voz humana. En la sección 6.3 se enumeran los problemas más

importantes asociados al reconocimiento automático de voz en un robot social. Seguidamente, en la sección 6.4 se describen los diferentes modos o paradigmas de realizar reconocimiento automático de voz. En la sección 6.5 se enumeran los requisitos a cumplir por el sistema de ASR. Posteriormente, en la sección 6.6 se realiza un estado del arte de los entornos de reconocimiento de voz disponibles. En la sección 6.7, se describe como integrar uno de estos entornos dentro de la arquitectura de control del robot. En la siguiente sección, 6.8, se describe como se ha integrado en la arquitectura dos motores de reconocimiento de manera concurrente. Le sigue, en la sección 6.9, un conjunto de experimentos realizados con el sistema de ASR. Se concluye, en la sección 6.10 con un resumen del capítulo.

6.2. Mecanismos de captura de sonido y voz humana

Recordar que la robustez y el alto rendimiento del ASR son objetivos primordiales, pero hasta ahora apenas se ha mencionado un aspecto crucial para el correcto desempeño de este tipo de interacción verbal: el sistema con el que se capturará el audio. Los robots deberían tener capacidades de audición similares a las del ser humano para llevar acabo una interacción natural en un contexto social. En entornos reales existen multitud de fuentes de ruido, por ello muchos robots tratan de evitar este problema forzando a los usuarios a interactuar usando un micrófono auricular colocado cerca de su propia boca. Para una interacción más natural, el robot debería escuchar los sonidos que le llegan mediante sus 'propios oídos" integrados en su cuerpo, en vez de forzar a los usuarios a llevar micrófonos auriculares inalámbricos [Breazeal, 2001].

Actualmente se puede agrupar los mecanismos de captura de audio en cinco tipos de micrófonos aplicables a la robótica, con independencia del conector usado (minijack, USB, bluetooth, WUSB, etc.). Cada uno de ellos ha sido diseñado para un tipo de uso, y para la interacción humano robot tienen determinadas ventajas e inconvenientes que se van a describir a continuación (ver Fig. 6.1):

(a) auricular (b) onminidireccional (c) array (d) Sobremesa (e) Vocal

Figura 6.1: Tipos posibles de micrófonos a usar en interacción humano-robot

- El *"auricular/headset"*, es un micrófono de tipo unidireccional, que es capaz de captar el sonido en un solo sentido y a escasa distancia de la boca del usuario. Este tipo de micrófono proporciona mayor calidad en el reconocimiento de voz y de robustez frente al ruido debido a esa cercanía directa con la voz del usuarios. Sin embargo, fuerza al usuario a llevar consigo el auricular y una "petaca" inalámbrica que actúa como emisor de la señal de audio. El robot necesita llevar incorporado la "petaca" receptora de la señal de audio emitida. Este tipo de micrófono es muy válido en diversas situaciones: cuando el ruido ambiente es muy alto, cuando la distancia de interacción es grande (debido a la trasmisión inalámbrica de la señal, la distancia de interacción no influye en la calidad de reconocimiento) y finalmente cuando existen varias personas conversando simultáneamente pero únicamente uno de ellos está interactuando con el robot.

- El micrófono de ambiente (omnidireccional), se caracteriza por su capacidad para captar sonido en todos las direcciones. Normalmente se emplea para captar el sonido de ambiente en eventos deportivos o en musicales. Su uso está muy limitado como dispositivo válido para el reconocimiento de voz, ya que el rango de frecuencias que recoge es muy amplio y rara vez tienen mecanismos de cancelación de ruido.

- El *"array de micrófonos"* es un número variable de micrófonos (unidireccionales) operando en paralelo colocados sobre una estructura fija. Se suelen montar en un número que varía entre 3 y 8 micrófonos. Su uso esta justificado por dos motivos: su capacidad de extraer con claridad la voz humana en ambientes ruidosos, sin necesidad de "petacas" ni auriculares; y por otro lado por poder desempeñar funciones de localización de la fuente sonora, es decir, determinar la dirección en la que viene el sonido y localizar al usuario respecto al robot. Estas características son muy interesantes en HRI y han sido estudiadas en varios trabajos con robots [Kim & Choi, 2007],[Tamai et al., 2005], [Valin et al., b], [Yoshida et al., 2009], [Tanaka et al., 2010]. Está creciendo su uso con videoconsolas [Chetty, 2009], portátiles, teléfonos móviles, coches[Oh et al., 1992], etc. Su principal inconveniente es su gran tamaño, debido a que los algoritmos de localización sonora y cancelación de ruido necesitan que los micrófonos estén suficientemente separados unos de otros. En la actualidad existen micrófonos de array comerciales disponibles, en esta tesis se han realizado pruebas con la captación de audio y localización sonora de los dispositivos Voice Tracker I [1] y la Microsoft Kinect [2].

- El *micrófono de sobremesa* es un tipo de micrófono unidireccional que es capaz

[1] http://www.acousticmagic.com/acoustic-magic-voice-tracker-i-array-microphone-product-details.html

[2] http://en.wikipedia.org/wiki/Kinect

de captar audio de una manera muy direccional a distancias que pueden oscilar entre los pocos centímetros y los varios metros. Han sido diseñados para tareas como videoconferencias en ordenadores de sobremesa y permiten al usuario conversar, sin necesidad de auriculares, sin por ello perder calidad de sonido. Son muy útiles para robots que no se pueden permitir la integración de elementos tan voluminosos como los arrays de micrófonos. La calidad de captura de voz es similar a estos últimos, si bien son algo menos robustos frente a condiciones de ruido ambiental.

- El *micrófono vocal* es usado por presentadores, cantantes, etc. Presenta varias características que hacen a estos dispositivos propicios para la interación humano robot. La primera de ellas, es que permite captar el rango de voz humana a una corta distancia, atenuando el resto de sonidos y voces más lejanas. La segunda, es que se puede encontrar modelos inalámbricos a un precio asequible (por ejemplo los micrófonos del juego *SingStar*). La tercera, es que puede facilitar la gestión del turno de palabra en una interacción con varios usuarios que se van intercambiando el micrófono.

6.3. Dificultades en la interacción por voz: *"Cocktail Party Problem"*

La captura de voz mediante micrófonos se encuentra con algunas dificultades que son necesarias de tener en cuenta para logar una correcta interacción, especialmente en el caso de micrófonos integrados en el propio robot. Existen multitud de fuentes de ruido que pueden degradar el reconocimiento de voz, como se puede ver en la figura 6.2.

6.3.1. Cocktail Party Problem

El "*Coctkail Party Problem*", literalmente traducido como el problema de la fiesta del cóctel, o de una forma menos literal, como el proceso de "atención selectiva", es el fenómeno por el que los humanos son capaces de focalizar la atención auditiva en un estímulo particular, mientras que filtran inconscientemente el resto de estímulos. El nombre viene del modo en que un asistente a una fiesta puede centrarse en una sola conversación en una sala ruidosa ([Haykin & Chen, 2005, Wood & Cowan, 1995, Bronkhorst, 2000, Conway et al., 2001]). Este efecto es el que permite a la mayoría de gente "sintonizar" con una sola voz y "desconectar" del resto. Esto también, ocurre en un fenómeno similar, el que ocurre cuando se detectan palabras de importancia originadas por un estímulo no atendido, por ejemplo, cuando se oye nuestro nombre en otra conversación. Ver Fig. 6.3.

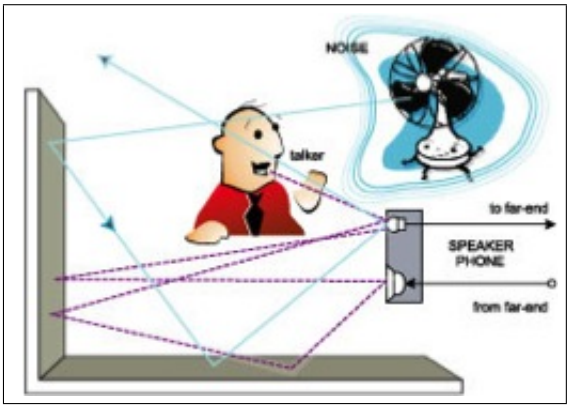


Figura 6.2: El robot capta muchos sonidos de su entorno, incluido el producido por él mismo al sintetizar voz. Es necesario filtrar todos esos sonidos no deseados

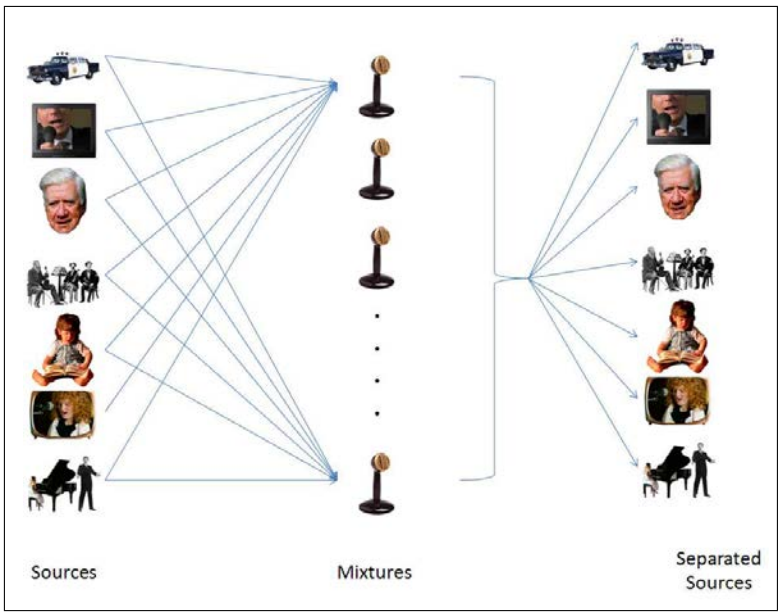


Figura 6.3: Cocktail Party Problem. Multitud de sonidos no deseados son captados por el robot. En necesario filtrar todas esas fuentes de sonido no deseadas, de manera equivalente al proceso de "atención selectiva" que somos capaces de llevar a cabo los humanos

6.3.2. Ruido de ambiente de fuentes diversas

Se puede hablar de ruido estacionario, o ruido blanco, al ruido constante y fácilmente predecible que producen elementos del entorno como son ventiladores, aires acondicionados, etc. Este tipo de ruido es el más sencillo de eliminar y al tipo de técnica que lo elimina se le denomina **"cancelación del ruido estacionario"**. Muchos micrófonos incorporan en su propio hardware filtros paso banda que eliminan la señal que queda fuera del rango de frecuencias de la voz humana y/o los componentes de señal con poca amplitud. También la mayoría de reconocedores de voz realizan esta tarea mediante software.

Existe otra fuente de ruidos que es la generada por el propio robot cuando se comunica. Es la propia voz del robot la que se puede acoplar en los micrófonos integrados en su cuerpo entorpeciendo el reconocimiento de voz. Si se desea una interacción natural, en la que el propio usuario pueda hablar al robot, inclusive mientras este último esté también hablando, es necesario de un mecanismo de **"cancelación activa del eco" (AEC)**. Multitud de desarrolladores no han tratado con este problema, por lo que simplemente tratan de evitarlo. Para ello, fuerzan al usuario a esperar a que el robot deje de hablar, sin tener capacidad de interrumpirle. Es decir, el reconocimiento de voz únicamente funciona por turnos, cuando el robot habla el ASR se encuentra desactivado (modo de interacción *half-duplex*). También, se colocan botones o conmutadores que permitan manualmente al usuario activar/bloquear la captura de audio. En este trabajo se ha tratado resolver el problema sin penalizar la naturalidad de la interacción, por ello se ha probado con diversos algoritmos software de cancelación activa de eco (*Ubuntu Pulseaudio Echo Canceller*), así como tarjetas de sonido con dicha funcionalidad incorporada.

En este trabajo se ha optado por una solución hardware, mediante la tarjeta de sonido USB *Creative Recon 3D*. Incorpora algoritmos de cancelación del eco, que se basan en restar (sumar con un desplazamiento de 180º en fase) la señal de audio generada y reproducida por los altavoces a la señal recibida por los micrófonos con algún tipo de transformación (ver Fig. 6.4). La diferencia entre la señal emitida por los altavoces y la recogida por el micrófono se da debido al tipo de micrófono, distancia entre ambos, etc, y necesita ser aprendida por el algoritmo de cancelación de eco. Normalmente transcurre un cierto tiempo, que suele ser de pocos segundos, en lo que el algoritmo "aprende" dicha transformación, en base a estimar unos parámetros.

Finalmente, existe un tercer tipo de fuente de sonidos, y son aquellos que no son ni estacionarios (fácilmente estimables) ni producidos por la propia voz del robot (de la que se conoce la señal), estos son sonidos producidos por los propios movimientos de las articulaciones del robot o de su base móvil, por fuentes externas como pueda ser el sonido del tren, u otros usuarios conversando cerca del robot. La eliminación de este tipo de ruido se suele llevar a cabo mediante el uso de micrófonos específicos para esta tarea (**"cancelación activa del ruido mediante microfono/s dedicado/s"**). El

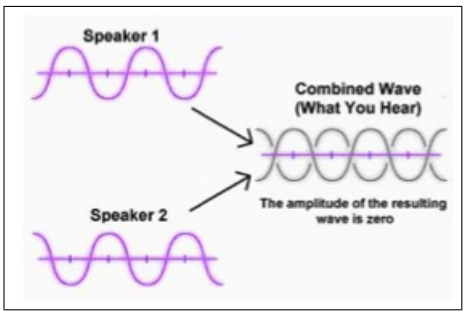


Figura 6.4: Cancelación activa del eco

funcionamiento de esta técnica se basa en restar la señal recibida por estos micrófonos dedicados a la señal recibida por el micrófono/s dedicados a captar la voz integrados en el robot.

Este tipo de técnicas, tanto la de cancelación activa de ruido como la de cancelación activa de eco, es frecuente de verlos integrados en teléfonos inteligentes (*smartphones*), ya que dedican uno o varios micrófonos para la captación de ruido, así como chips integrados para *DSP*, que procesan estas señales para la cancelación del ruido y de la propia voz del otro interlocutor en modo "manos libres". Sin embargo, en un robot social, es tremendamente complicado lograr una configuración de micrófonos que permitan captar la fuente de ruidos y que no capten la voz del usuario. Si la presencia de ruidos difícilmente estimables es muy alta, la interacción se ve abocada a interactuar mediante micrófonos auriculares inalámbricos, que son los mas adecuados para estas circunstancias. Además, esta fuente de ruidos no estacionarios, a parte de entorpecer el reconocimiento de voz, también dificultan las tareas de localización de la fuente sonora. En trabajos como [Qi, 2008], se están introduciendo algoritmos y técnicas software para tratar el problema de la *"cancelación activa de ruido"* o *"control activo del ruido"*.

Existe otro campo de investigación relacionado con el control activo del ruido y el reconocimiento de voz: *"Voice Activity Detector" (VAD)*. Trata de determinar cuando empieza y finaliza la voz humana. Tradicionalmente, se ha computado mediante el establecimiento de un umbral de volumen. Una vez que la señal de audio recibida supera dicho umbral de volumen, se considera el comienzo de la locución. Se considera el fin de la locución, cuando no se supera el umbral establecido durante un instante de tiempo determinado (normalmente 800 milisegundos). Esta aproximación presenta el problema de que un ruido fuerte, por ejemplo, un sonido de un animal, una canción, etc. pueden ser considerados voz humana. En la literatura se puede encontrar varias trabajos que tratan el tema [Barrett, 2000, Freeman & Boyd, 1993, Barrett, 1998, Freeman & Boyd, 1994], y que plantean sistemas más avanzados. Alguno de estos sistemas consiguen en entornos reales, como en un coche circulando, filtrar la voz del

conductor, suprimiendo los sonidos de fondo, como los de la radio, el motor, etc.

En nuestro caso, se ha desarrollado un detector de voz que tiene en cuenta varios descriptores de la señal, en tres dominios de la señal diferentes [Alonso-Martin et al., 2013].

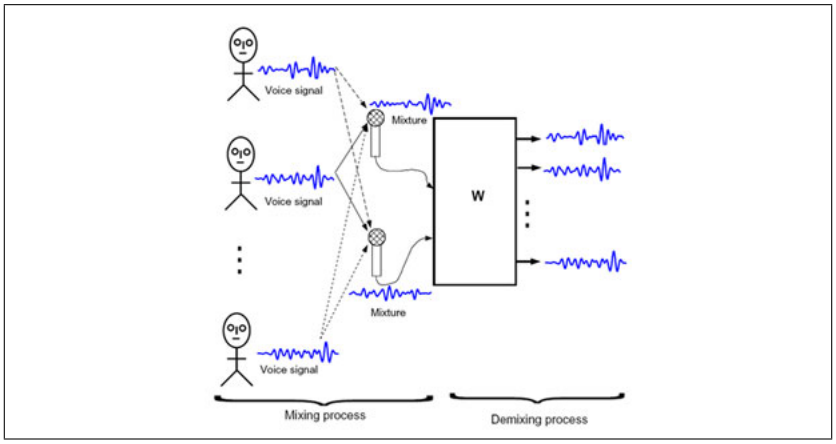


Figura 6.5: Mezcla a la entrada y separación a la salida de los canales de audio

Además de los problemas relacionados con el ruido, existe la problemática de que pueden existir varios usuarios hablando simultáneamente entre ellos y/o con el robot (recordar el "*coctkail party problem*" y la Fig. 6.5). Si la interacción está basada en micrófonos auriculares, la separación de dichos canales es trivial, puesto que cada uno de ellos usa un dispositivo de audio diferente, y por lo tanto un canal independiente. En cambio, si la interacción se realiza mediante micrófonos integrados en el propio robot, las locuciones de cada usuario se reciben entremezcladas en un sólo canal. Para un correcto reconocimiento de la voz de cada usuario, existen técnicas de **"separación de la fuente sonora" (SSS)**. El paquete de ROS, denominado HARK, incluye entre otras funciones dicha SSS. Ejemplos de SSS aplicados a la música se pueden ver en: http://bass-db.gforge.inria.fr/fasst/, en este sitio se puede descargar un software en Matlab que posibilita diferenciar de una canción cada linea instrumental.

6.4. Paradigmas en el reconocimiento automático de voz

Como se ha dicho en la introducción, hay dos tipos de reconocedores de voz. Los que funcionan para cualquier usuario ("**independiente del hablante**"), y los que únicamente funcionan con el usuario para el que se ha entrenado el modelo ("**dependiente del hablante**"). Estos últimos, necesitan de un entrenamiento del reconocedor

con el usuario, que se realiza leyendo determinados textos. Por razones de naturalizad en la interacción este trabajo se centra únicamente en los sistemas que funcionan con independencia del locutor.

Dentro de estos "sistemas independientes del hablante", existen varios métodos o paradigmas de llevar a cabo el reconocimiento automático de voz. Antes de explicar cada uno de ellos es necesario entender que dentro del proceso de reconocimiento de voz hay que distinguir dos fases: 1) transcribir la frase pronunciada verbalmente a texto y 2) entender el significado de lo que se ha reconocido (extracción de los valores semánticos). La segunda parte es tan importante o más que la primera, puesto que los gestores de diálogo (especialmente los basados en rellenar huecos de información) reciben y trabajan sobre la información semántica. Teniendo claras estas dos fases se verán los modos de reconocimiento de voz posibles.

6.4.1. Basado en gramáticas

Es el método más habitual, consiste en establecer un conjunto de reglas gramaticales. Estas reglas restringen considerablemente el conjunto de oraciones lingüísticamente válidas para el contexto comunicativo en el que se realiza la interacción. Este conjunto de reglas gramaticales, corresponde a lo que se conoce como gramáticas de "contexto libre", en la que cada regla es de la forma: $V ==> W$. Donde V es un símbolo no terminal y W es una cadena de terminales y/o no terminales. El término *libre de contexto* se refiere al hecho de que el no terminal V puede siempre ser sustituido por W sin tener en cuenta el contexto en el que ocurra.

En el reconocimiento basado en gramáticas, el reconocedor sólo es sensible a un subconjunto del lenguaje natural, con la consiguiente perdida de naturalidad en la interacción. Sin embargo, aumenta la precisión de reconocimiento. Suele ser útil para diálogos en el que el conjunto de opciones disponibles es limitado. Otra ventaja fundamental, a parte de su alta precisión para contextos limitados, es que permite fácilmente extraer el significado semántico de la frase reconocida mediante el propio uso de las gramáticas, es decir, las reglas semánticas se introducen conjuntamente con las reglas gramaticales en un mismo fichero. Este tipo de reconocedor de voz es ampliamente usado en ciertas aplicaciones telefónicas en el que el conjunto de opciones son limitadas (darse de alta, baja, consultar facturas...); para reservar billetes de avión, trenes; también para interactuar con un robot a modo de asistente del humano ("leeme las noticias","preséntate",'enciende la TV" ...). Herramientas como Loquendo ASR, Nuance Recognizer, y Sphinx, implementan este tipo de reconocimiento de voz.

Para ver ejemplos concretos de gramáticas construidas para este trabajo consultar el apéndice de este documento.

6.4.2. Basado en un modelo estadístico del idioma (SLM)

Este método permite reconocer texto libre sin restringir el lenguaje natural a un subconjunto del mismo (como sí lo hacen los basados en gramáticas). Dentro de estos modelos estadísticos, existen los llamados 2-gram, 3-gram, 4-gram ... que indican la probabilidad de que aparezca en la frase una palabra habiendo aparecido justo anteriormente las "n" palabras anteriores. En ese sentido, un modelo 4-gram tendrá en cuenta las 4 palabras anteriores para determinar la quinta palabra. Para construir estos modelos estadísticos se necesita un enorme corpus con ejemplos de diálogos en dicho idioma. Su principal ventaja es que la interacción no se encuentra restringida a ciertas frases, por contra sus desventajas son la importante perdida de precisión de reconocimiento y la necesidad de un módulo adicional que realice la comprensión semántica de la frase, proceso conocido como Natural Language Understanding (NLU). Este tipo de reconocimiento es muy útil en diálogos en los que se desea obtener información sobre cualquier cosa, como por ej: "hola, ¿me podrías decir quién es Rafael Nadal?", o "me gustaría escuchar una canción de Mecano". Un reconocedor basado en un modelo estadístico del lenguaje, 5-gram, es Google ASR.

6.4.3. Basado en un modelo estadístico de un contexto específico

Una manera de aumentar la precisión de reconocimiento del modelo estadístico es restringir en cierta medida el modelo del lenguaje. De esta manera, esta se encuentra a medio camino de las dos anteriores. El modelo estadístico se entrena únicamente con las posibles frases que se usan en interacciones reales con el robot y no con el conjunto de posibles frases que se pueden decir en el idioma escogido. Esto presenta una restricción en el conjunto de las interacciones posibles pero un aumento importante de la precisión de reconocimiento. Por otro lado, sigue siendo necesario un módulo adicional que realice la tarea de extracción de valores semánticos del texto reconocido (NLU). Este tipo de reconocimiento es muy útil en diálogos de pregunta abierta, en el que el conjunto de opciones es bastante elevado. Por ejemplo el robot puede preguntar al usuario: "¿qué deseas hacer?" (modo pregunta abierta), en el cuál el conjunto de opciones posibles es en cierto modo limitado por el contexto en el que se produce. Existen herramientas que facilitan la tarea de construir este modelo estadístico del contexto, como es Nuance SayAnything.

6.5. Requisitos para el sistema de reconocimiento de voz efectivo

En nuestra arquitectura de control, la capacidad de reconocimiento automático del habla debe ser implementada e integrada en la arquitectura como una "habilidad". Esto permite que cualquier componente de la arquitectura pueda usar la funcionalidad de reconocimiento de voz fácilmente. Para lograr este objetivo se han definido los requisitos necesarios para lograr un actualizado y potente sistema de ASR:

- **Independencia del hablante**: el sistema debe reconocer el lenguaje natural con independencia de quién es la persona que esté hablando y si está tiene o no entrenamiento previo con el reconocedor. Por lo tanto debe funcionar para cualquier usuario, sea hombre o mujer, niño, adulto o anciano, sea experto en tecnología o la primera vez que interactúa con el robot.

- **Alta precisión de reconocimiento de voz**: los resultados del reconocimiento deben ser tan precisos como sea posible. En un caso ideal, el reconocimiento debería tener una precisión similar a la que tiene un humano de reconocer el lenguaje hablado expresado por otro humano.

- **Soporte para micrófono**: normalmente los sistemas de reconocimiento de voz son diseñados para aplicaciones telefónicas, sin embargo se necesita obtener el audio mediante micrófonos conectados al robot (computador). El micrófono debe estar continuamente enviando muestras de audio al reconocedor de voz. En aplicaciones telefónicas las muestras de audio suelen ser enviadas desde un servidor al reconocedor de voz mediante ficheros.

- **Soporte para Linux**: dado que los robot en los que se usa el sistema usan un sistema operativo Linux, se necesita que el sistema de reconocimiento de voz sea compatible con este sistema y con nuestra configuración software/hardware.

- **Cambio de gramáticas en tiempo real**: el reconocimiento de voz está basado en gramáticas de contexto libre. Con el uso de gramáticas se restringe considerablemente las opciones comunicativas, mejorando la precisión del reconocimiento de voz, pero con el inconveniente de perder naturalidad en la interacción. Debería ser posible la carga/descarga de una o varias gramáticas durante la ejecución del diálogo sin la necesidad de detenerse.

- **Detector de habla**: el sistema debe ser capaz de distinguir la voz respecto al silencio, un ruido puntual, o ruido de fondo. Para ello, el sistema debe tener mecanismos de cancelación de ruido y un umbral para diferenciar silencio de voz. Para lograr esto, suele ser necesaria de una combinación de mecanismos

software y hardware. Con un sistema adecuado de detección de habla no es necesario del uso de un botón, o cualquier otro mecanismo mediante el cual el usuario indicaría al robot cuando empieza y cuando finaliza de hablar.

- **Soporte a gramáticas semánticas**: las gramáticas semánticas facilitan la tarea de extraer el significado y la información relevante para el dialogo. Las gramáticas semánticas añaden a las gramáticas formales un post-procesado de la información reconocida, mediante un lenguaje de script que incorporado en la propia definición de la gramática formal, permite la extracción de los valores semánticos.

- **Soporte de estándares**: en tecnologías de voz y del habla se han definido varios estándares que resultan útiles para acelerar el proceso de desarrollo de aplicaciones de voz. Es altamente recomendable ceñirnos a estos estándares. Los mas importantes en ASR son:

  - SRGS: Speech Recognition Grammar Specification
  - NLSML: Natural Language Semantics Markup Language
  - SISR: Semantic Interpretation for Speech Recognition

- **Alta eficiencia**: la posibilidad de usar el reconocimiento de voz con bajo consumo de energía y CPU es deseable, puesto que se va a ejecutar conjuntamente con el resto de módulos de la arquitectura de control del robot.

- **Soporte multilenguaje**: debido a que el sistema de interacción permite la comunicación en varios idiomas, es necesario que el reconocedor de voz pueda entender y trabajar con gramáticas construidas en diversos idiomas. En nuestro caso, se ha trabajado con español, inglés americano e inglés británico, y es fácilmente ampliable a cualquier otro idioma.

- **Identificación del hablante mediante la voz**: dado que el sistema de diálogo desarrollado necesita identificar a los usuarios para cargar su perfil de usuario y así personalizar el diálogo a cada interlocutor, se hace necesario que el reconocedor de voz sea capaz de reconocer no sólo la frase pronunciada sino identificar también que usuario está hablando (de entre el conjunto de posibles).

- **Adaptabilidad del modelo acústico**: normalmente el motor de reconocimiento de voz está entrenado para ambientes telefónicos de atención al ciente (call cliente), por lo que puede ser deseable reentrenar el modelo acústico a nuestras verdaderas necesidades para incrementar la precisión del reconocimiento.

- **Capaz de reconocer texto libre (modo dictado)**: algunas veces se requiere interactuar con el robot en modo "dictado" o "lenguaje abierto", es decir, sin la restricción de gramáticas. En este caso, es necesario del uso de modelos estadísticos del lenguaje escogido. Para la construcción del modelo, es necesario un enorme corpus lingüístico. El uso de este tipo especial de reconocimiento de voz, es muy útil para ciertos momentos, en que el número de posibilidades comunicativas es tan grande que es difícil recoger todas ellas en una gramática. Por ejemplo: grupos musicales, nombres, apellidos, etc.

6.6. Análisis de los entornos de reconocimiento de voz disponibles

Una vez que se han definido los requisitos que debe tener nuestra habilidad de reconocimiento, se necesita estudiar qué soluciones software se adecuan mejor a tales requerimientos. Basándonos en la literatura [Kibria & Hellström, 2007] y en nuestra propia experiencia, se han seleccionado los sistemas de reconocimiento más conocidos y a priori los más potentes. Se ha realizado estudio pormenorizado sobre cada uno de ellos, comparando unos con otros, en base a los requisitos definidos en la sección anterior.

Finalmente, los cinco sistemas sometidos a estudio comparativo han sido los siguientes:

- Verbio ASR v8[3] (Windows).
- Nuance Recognizer v9[4] (Linux).
- Nuance VoCon 3200[5](Windows).
- Loquendo ASR v7.7[6] (Windows).
- Sphinx IV[7]

De todos ellos se consiguió una licencia de prueba. El estudio consistió, para cada uno de ellos, en el análisis de la documentación asociada, su configuración para nuestro

[3] `http://www.verbio.com/`
[4] `http://www.verbio.com/`
[5] `http://spain.nuance.com/vocon/`
[6] `http://www.loquendo.com/es/technology`
[7] `http://cmusphinx.sourceforge.net`

entorno de trabajo, y la ejecución de una batería de pruebas para analizar cada uno de los requisitos propuestos [8]

La mayor parte del estudio que se ha realizado se resume en la Tabla 6.6. En ella, cada columna representa un framework de reconocimiento distinto y cada fila el valor que toma para cada uno de los requisitos analizados. La mayoría de las características a estudio se han determinado de una manera objetiva, por ejemplo determinar si un producto tiene o no "speech detector", en cambio otras características, como la usabilidad del producto, es fruto de nuestra opinión subjetiva en el uso de dicha herramienta.

Basándonos en este estudio y en nuestras necesidades para mejorar la interacción humano-robot, finalmente se ha decidido elegir el entorno software Loquendo, para su integración dentro de nuestra arquitectura de control. Su elección se justifica debido a que Loquendo ASR cumple con todos los requerimientos descritos previamente. Otro aspecto decisivo en su elección ha sido su buena relación calidad/precio.

El resto de productos han sido descartados debido a las siguientes razones:

- Nuance Recognizer v9: aunque es un framework con una buena tasa de acierto, está diseñado para aplicaciones telefónicas y hasta la fecha, no da soporte a la entrada de audio directamente desde el micrófono. Además su coste es bastante elevado.

- Nuance Vocon: es otro producto de Nuance, con buen rendimiento, pero hasta la fecha no tiene versión Linux disponible.

- Verbio ASR v8: es el de menor precisión de entre los estudiados, además no es capaz de proporcionar resultados parciales de reconocimiento. Es decir, hasta que no finaliza completamente la oración el interlocutor y procesa en su totalidad dicha frase, no es capaz de ir proporcionando resultados de reconocimiento.

- Sphinx IV: es un framework de software libre, con soporte a multitud de idiomas y sistemas operativos. Ha sido desarrollado por la universidad Carnegie Mellon, Sun Microsystems, Hewlett Packard, entre otros, pero su mayor problema es que tiene menos precisión de reconocimiento que Nuance y Loquendo.

[8] Resaltar, que no ha sido una tarea sencilla la obtención de las licencias, la lectura de grandes cantidades de documentación, así como la tarea de tener configurado y funcionando cada sistema en un entorno real de trabajo; si bien todas esas fases han sido necesarias para poder hacer un riguroso análisis comparativo de las mismas.

	Verbio ASR	Nuance Recognizer V9	Nuance VoCon	Loquendo ASR	Sphinx IV
País Desarrollo	España	USA	USA	Italia	USA
Sin entrenamiento	Sí	Sí	Sí	Sí	No
Independiente del interlocutor	Sí	Sí	Sí	Sí	Sí
Basado en gramáticas	Sí	Sí	Sí	Sí	Sí
Modelo estadístico del lenguaje	Sí	Sí	No	Sí	Sí
Sistema Operativo soportado	Linux / Windows	Linux / Windows	Windows	Linux / Windows	Linux / Windows / Solaris
Detección del habla	Sí	Sí	Sí	Sí	Sí
Soporte PC micrófono	Sí	No	Sí	Sí (pero con parche)	Sí
Multilenguaje	Sí	Sí	Sí	Sí	Sí
Basado en fonemas	Sí	Sí	Sí	Sí	Sí
Identificador del hablante	No	No	No	Sí	No
Modo "WordSpotting"	Sí	Sí	No	No	No
Gramáticas semánticas	No	No	Sí	Sí	No
Aprendizaje de nuevas palabras	No	No	Sí	No	No
Adaptación del modelo acústico	No	Sí	No	Sí	Sí
Usabilidad	Alta	Baja	Muy alta	Baja	Media
Ejemplos de uso	Suficientes	Suficientes	Muy buenos	Suficientes	Buenos
Recursos adicionales	Pobre	Suficientes	Muy buenos	Buenos	Normal
Soporte	Sí	Sí	Sí	Sí	No
Facilidad de compra	Fácil	Difícil	Difícil	Difícil	Muy sencilla
Precisión	Media	Alta	Alta	Alta	Media
Precio	Bajo	Alto	Medio-Bajo	Medio	Gratuito

Cuadro 6.1: Comparación de los principales entornos de reconocimiento de voz (2011)

6.7. Integración con la arquitectura de control

Una vez que se ha elegido el producto que mejor se adapta a nuestras necesidades, se debe integrar con la arquitectura de control. Como se ha dicho, en la arquitectura de control AD, se denomina "habilidad" a cualquier componente de software que proporciona una nueva "capacidad" al robot. El objetivo ha sido integrar la funcionalidad de reconocimiento automático del habla como una "habilidad" dentro de esta arquitectura AD.

La habilidad que se ha desarrollado está estructurada en tres capas de abstracción. En el nivel inferior se encuentra el motor de reconocimiento, que está compuesto por las librerías proporcionadas por Loquendo. Sobre estas librerías, que son las que proporcionan las funcionalidades de reconocimiento propiamente dichas, se han escrito algunas funciones básicas, a las que se han denominado "ASR primitivas". Estas primitivas implementan las mas importantes funcionalidades necesarias, como son establecer gramáticas, arrancar y parar el reconocimiento de voz, obtener los resultados del reconocimiento de voz, establecer el formato de la entrada de audio, etc.

Finalmente, sobre estas primitivas de voz se ha construido la habilidad propiamente dicha, "ASRSkill". Esta habilidad se ajusta a la misma plantilla que el resto de habilidades que forman parte de la arquitectura de control. Cualquier habilidad que quiera realizar funciones de reconocimiento de voz, utiliza la funcionalidad que ofrece la habilidad que se acaba de describir.

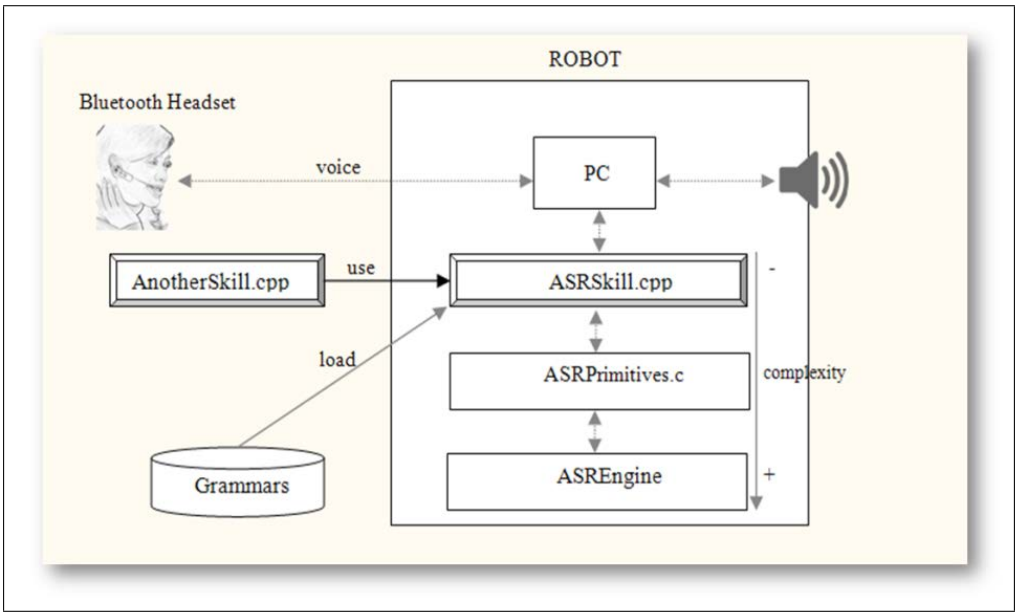


Figura 6.6: Arquitectura del sistema de reconocimiento de voz

Esta estructura en capas, se puede ver en la Fig. 6.6, se corresponden con los tres niveles de abstracción del Procesamiento del Lenguaje natural:

1º Nivel -léxico-: Reconocimiento de voz

- Modelos acústicos del lenguaje
- Gramáticas
- ¿Qué ha dicho el hablante?

2º Nivel -semántico-: Análisis de voz

- Significado léxico-gramatical
- ¿Qué quiere decir lo que ha dicho?

3º Nivel -pragmático-: Entendimiento

- Contexto del discurso – conocimiento sobre el dominio del discurso
- What has the caller asked?

El primer nivel (reconocimiento del habla) se relaciona con el motor de reconocimiento proporcionado por Loquendo. El segundo nivel se relaciona con las primitivas de voz, mientras que finalmente el tercer nivel casa con la habilidad propiamente dicha. Finalmente sobre todas ellas se sitúa el "gestor del dialogo" que gestiona la interacción usando los diversos módulos: ASR, TTS, Localización Sonora, Identificación del hablante, etc.

6.8. Motores de reconocimiento concurrentes

Normalmente, los programadores de aplicaciones de voz usan el valor de confianza para decidir si aceptar o rechazar los resultados arrojados por el reconocedor de voz. Para ello, es necesario fijar un umbral, de tal manera que si se supera se considera reconocimiento correcto y si no se supera se considera fallido. Existen sistemas de diálogo basados en modelos estadísticos, como los POMDP, que mantienen simultáneamente varias hipótesis de los reconocimientos realizados (n-Best), pero no está claro si en la práctica estos sistemas pueden ser usados eficientemente, debido a que su carga computacional crece exponencialmente con el tamaño del problema [Young et al., 2010, Young, 2006, Williams & Young, 2007, Minami et al., 2010, Zhou & Yuan, 2010, Roy et al., 2000, Roy et al., 1998].

Como se acaba de comentar, la mayoría de los sistemas de diálogo están basados en el uso local del valor de confianza y de un umbral general. Algunas mejoras permiten

adaptar dicho umbral dinámicamente o cambiar de estrategia, mediante el uso de preguntas implícitas y/o explícitas.

En esta tesis se propone e implementa un nuevo modo de usar los resultados del reconocimiento de voz para clasificarlos como válido o inválidos. Esta nueva vía sigue usando los valores de confianza, pero de una manera mas inteligente. Nosotros lo llamamos mecanismo de "segunda opinión". Para ello, se ha desarrollado un nuevo componente que usa diferentes fuentes de conocimiento para calcular un nuevo valor de confianza mas fidedigno.

"Segunda opinión" está basado en el uso de varios motores de reconocimiento de voz corriendo concurrentemente (al menos dos). En este trabajo, se han usado dos motores en paralelo, el primero de ellos Loquendo ASR, descrito anteriormente y el segundo motor de reconocimiento ha sido Google ASR (se trata de un servicio web online que actualmente es usado por Google en Google Chrome, Android, Youtube Automatic Subtitles, etc), pero que haciendo "ingeniería inversa" se ha podido integrar satisfactoriamente en nuestro sistema.

Los resultados arrojados por Loquendo y Google ASR pueden ser procesados por el módulo que computa un nuevo valor de confianza. Para calcular ese nuevo valor, se ha usado la ecuación 6.1.

C1 = valor confianza ASR1(es proporcionado en cada reconocimiento por el primer motor de reconocimiento y es usado para indicar la garantía que tiene el reconocedor de que el resultado es el correcto).

C2 = valor confianza ASR2 (es proporcionado en cada reconocimiento por el segundo motor de reconocimiento y es usado para indicar la garantía que tiene el reconocedor de que el resultado es el correcto).

SNR1 = Proporción de Señal a Ruido para ASR1(es proporcionado para cada reconocimiento por el primer motor de ASR y es indicativo de la calidad de señal recibida).

SNR2 = Proporción de Señal a Ruido para ASR2 (es proporcionado para cada reconocimiento por el segundo motor de ASR y es indicativo de la calidad de señal recibida).

SR1 = pre-calculada tasa de acierto para el ASR1 (tasa de acierto obtenida empíricamente después de realizar test en entornos reales con el primer reconocedor, indica el porcentaje de veces que el primer reconocedor clasifica correctamente los reconocimientos, acertando).

SR2 = pre-calculada tasa de acierto para el ASR2 (tasa de acierto obtenida empíricamente después de realizar test en entornos reales con el segundo reconocedor,

indica el porcentaje de veces que el segundo reconocedor clasifica correctamente los reconocimientos, acertando).

AC = confianza media final (confianza resultante de tener en cuenta los resultados obtenidos por cada reconocedor y su relativo peso en el cálculo de esta confianza media).

$$AC = \frac{SR1}{SR1 + SR2} * C1 + \frac{SR2}{SR1 + SR2} * C2 \tag{6.1}$$

Con esta función (6.1) se da más valor/peso al reconocedor que tiene más tasa de acierto "a priori", pero también se tiene en cuenta el segundo reconocedor, con lo que se da mayor fiabilidad a los resultados ofrecidos. Se ha calculado la tasa de acierto de los distintos reconocedores probados y se muestra en la tabla 6.6.

En la ecuación (6.1) no se ha tenido en cuenta la tasa de acierto de los reconocedores de voz según el nivel de ruido del entorno. Como se ha visto el reconocedor de voz está altamente influenciado por el ruido, por lo tanto la tasa de acierto no debería ser una función uniforme para todos los valores de ruido. Probablemente unos reconocedores de voz están mas afectados por el ruido de ambiente que otros. Por esa razón parece lógico penalizar los más afectados por el ruido en condiciones adversas y recompensarlos en condiciones de silencio. Teniendo en cuenta esta consideración, se ha reformulado la función anterior, y la nueva ecuación (Eq. 6.2) relaciona la calidad voz/ruido (SNR) con la tasas de acierto (ver Fig.ResumenHeadSets).

El valor de la relación voz/ruido (calidad del audio que llega al robot libre de ruidos) es arrojado por cada reconocedor en cada reconocimiento de voz y se puede, en base a ese valor y a la tasa de acierto que se va acumulando para cada intervalo de SNR, construir una función que se llama PSR para cada uno de los reconocedores usados (ver Fig. 6.7).

Recordar que el valor de ruido arrojado por SNR está muy influenciado por el tipo y modelo de micrófono, el nivel de reverberación de la sala, por el tipo de ruido, y por supuesto por el volumen del mismo.

$$AC = \frac{PSR1(SNR1)}{PSR1(SNR1) + PSR2(SNR1)} * C1 + \frac{PSR2(SNR2)}{PSR1(SNR2) + PSR2(SNR2)} * C2 \tag{6.2}$$

En la Fig. 6.8 se compara la confianza-media calculada usando la ecuación formulada anteriormente (6.2) y la confianza "tradicional" arrojada por uno sólo de los reconocedores, en este caso Loquendo. Los dos ejes de la gráfica son "falsas alarmas" y "ausencias" que se definen en las ecuaciones 6.3 y 6.4, con ayuda de la tabla 6.2.Básicamente en el eje X se representa las frases aceptadas que no deberían de serlo y en el eje Y se representa las frases rechazadas que eran correctas.

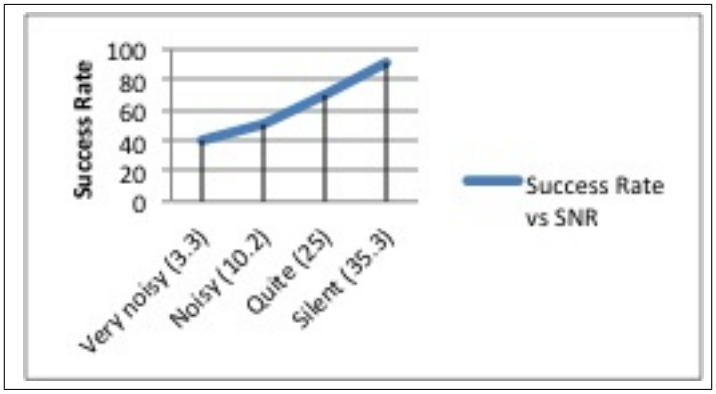


Figura 6.7: Tasa de acierto vs SNR: función de probabilidad de tasa de acierto (PSR)

	Predicho como positivo	Predicho como negativo
Realmente positivo	Verdadero Positivo (VP)	Falso Negativo (FN)
Realmente negativo	Falso Positivo (FP)	Verdadero Negativo (VN)

Cuadro 6.2: Tabla de definiciones

$$Falsas_a larmas = \frac{FP}{FP + VN} \tag{6.3}$$

$$Ausencias = \frac{FN}{VP + FN} \tag{6.4}$$

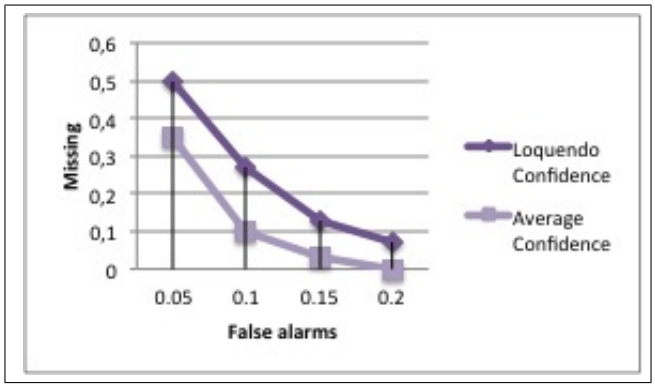


Figura 6.8: Confianza media vs típica confianza

6.9. Experimentos sobre la precisión del reconocimiento

Ahora que se ha elegido un framework y que se ha integrado en nuestra arquitectura de control, se necesita probar la habilidad implementada en entornos reales y con diferentes tipos de micrófonos. Para ello, se han diseñado varios escenarios de prueba con el objetivo de analizar la precisión del reconocedor dependiendo de: ruido de ambiente, volumen del hablante, entonación, edad, género y el tipo de micrófono usado.

Cada test ha sido llevado a cabo con diferentes usuarios diciendo diferentes frases, usando la misma gramática y sin entrenamiento previo con el sistema. En los primeros escenarios de test, se ha realizado la adquisición de audio usando micrófonos profesionales unidireccionales e inalámbricos. El micrófono debe estar situado a pocos centímetros de la boca del usuario y mirando hacia dicha boca (recordar que el micrófono es unidireccional, por lo que sólo capta sonido en un sentido). La señal acústica recibida por el micrófono es transmitida inalámbricamente al receptor integrado en el robot. Tanto el emisor como el receptor utilizados[9] se pueden ver en la Fig 6.9.

Figura 6.9: Receptor-emisor: micrófonos inalámbricos Sennheiser

Los usuarios comenzaban a decir frases válidas para la gramática establecida en el test. Ellos decían frases continuamente, con una ligera pausa entre cada frase, sin entrenamiento previo con el sistema, pero sí con el conjunto de posibles frases y construcciones válidas para dicha prueba. Mientras tanto, se monitorizaba el experimento y se tomaba nota sobre los resultados arrojados por cada reconocimiento de voz.

[9] El modelo específico de transmisor usado puede encontrase en: `http://www.sennheiser.com/sennheiser/home_es.nsf/root/professional_wireless-microphone-systems_broadcast-eng-film_ew-100-series_021418` El modelo específico de micrófono auricular puede verse en: `http://www.sennheiser.com/sennheiser/home_es.nsf/root/professional_wireless-microphone-systems_headsets_headsets_009862` (500$) y `http://www.logitech.com/en-us/webcam-communications/internet-headsets-phones/devices/3621` (40$)

Una vez que el usuario completaba su turno de reconocimientos de voz, después de decir 100 oraciones, se pasa a repetir la misma prueba con el siguiente usuario, hasta completar un total de 10 usuarios. Por lo tanto, cada uno de los escenarios que se han presentado ha sido probado con un total de mil frases.

Para estimar la precisión del reconocimiento, se han analizado dos parámetros: la tasa de acierto y el valor de confianza. La tasa de acierto hace referencia al porcentaje de veces que el reconocedor es capaz de reconocer correctamente la frase pronunciada, mientras que el valor de confianza hace referencia a la garantía o estimación que hace el propio reconocedor de voz de que el resultado que está ofreciendo es el correcto. Si el valor de confianza está muy cerca de 1, el reconocedor de voz esta casi seguro que la frase reconocida coincide con la elocución del usuario, mientras que valores próximos a 0, indican que hay poca confianza en que el resultado ofrecido tenga que ver con lo que el usuario dijo.

La tasa de acierto puede ser usada para comparar la precisión de reconocimiento entre diferentes motores o software de reconocimiento de voz. Dicha tasa de acierto rara vez es proporcionada por los comerciales (al menos no la tasa real de acierto) y es necesario de pruebas en entornos reales para acercarnos a conocer la precisión real de los mismos.

6.9.1. Utilizando micrófonos auriculares inalámbricos

Sin ruido importante de fondo

En este primer escenario los usuarios se comunican con el robot en un entorno cerrado como es el laboratorio, sin ruido de ambiente significativo. El rango de sonido ambiente aproximado se situó entre 40 y 45 dB. Este tipo de ruido es conocido como ruido estacionario y es producido por el propio robot, por los ordenadores, ventiladores, etc. Este tipo de ruido es relativamente sencillo de estimar y de eliminar.

Los resultados obtenidos nos dan un valor de confianza media de 0.722 siendo 0 el mínimo valor y 1 el máximo. En el 99 por ciento de los casos, las oraciones pronunciadas fueron correctamente reconocidas y transcritas. Recordar que las frases pronunciadas casan perfectamente con la gramática establecida en el test.

Con estos resultados, se puede concluir que la precisión del reconocedor de voz con varios usuarios, en condiciones de relativo silencio, es bastante elevada (usando este tipo de micrófonos). El valor de confianza es alto y la tasa de acierto es prácticamente del 100 por ciento. Estos resultados experimentales se ajustan a los resultados oficiales proporcionados por Loquendo en [Paolo Baggia, 2005] y [Dalmasso et al.,].

Con ruido de fondo

Este escenario es muy similar al descrito anteriormente, exceptuando que se añade ruido de fondo. Para ello, se dejó una televisión encendida con un canal musical. La televisión se encontraba a un distancia de 7 metros del usuario, a un volumen de entre 65 y 70 dB (medido en la posición del usuario).

Los resultados obtenidos en este test son un 0.703 valor de confianza y un 98 por ciento la tasa de acierto. Estos valores están cercanos a los del test anterior, por lo que se puede seguir considerando un reconocimiento de voz preciso bajo estas condiciones de ruido ambiental y micrófonos auriculares.

Es muy importante que el micrófono esté colocado muy cerca de la boca del usuario, ya que el micrófono es unidireccional (capta sólo el sonido que en un único sentido y con poco alcance). Este tipo de micrófonos son comúnmente usados en cualquier teléfono inteligente actual. Además, hay que tener en cuenta que los reconocedores de voz han sido entrenados para trabajar especialmente con este tipo de micrófonos.

Con ruido cerca del usuario

En este nuevo escenario se ha colocado la fuente de ruido muy cerca del usuario, para ellos se ha puesto a sonar música por los altavoces a menos de un metro del usuario. El rango de sonido emitido varía entre 70 y 75 dB, lo que se puede considerar un volumen bastante elevado.

Los resultados obtenidos son de un valor medio de confianza de 0.673 y una tasa de acierto del 97 por ciento. Estos valores vienen a confirmar que el reconocedor de voz, con la combinación de micrófonos auriculares, es robusto frente a condiciones de ruido adversas. Una de las principales razones de seguir obteniendo estas tasas de acierto tan elevadas es que se sigue obteniendo una buena relación entre señal de voz frente a la señal generada por el ruido (SNR), debido a la propia naturaleza del micrófono. Otro motivo es que el reconocedor ha sido entrenado bajo condiciones de ruido. Además, el propio motor de reconocimiento es capaz de cancelar el ruido estacionario.

En recientes estudios realizados, se ha mejorado el proceso de adquisición de voz con la adquisición de una tarjeta de sonido que incorpora cancelación de ruido y cancelación de eco. La cancelación de ruido se logra mediante filtros paso banda, mientras que la cancelación activa de eco consigue eliminar de la propia entrada de audio la voz emitida por el propio robot cuando está hablando, evitando su acople a la señal de voz del usuario.

Se puede concluir que en ambientes ruidosos, incluso en condiciones realmente adversas, ASRSkill es muy robusta y precisa usando la configuración hardware apropiada (micrófonos y tarjeta de sonido adecuada). Ver la Fig. 6.10 a modo de resumen.

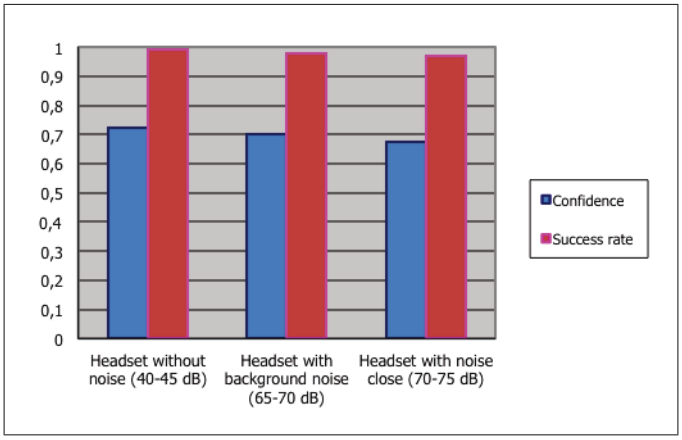


Figura 6.10: Resumen de la precisión de reconocimiento usando micrófonos auriculares

Diferentes volúmenes de voz

En este escenario se ha analizado como influye el volumen de la voz del usuario en la precisión del reconocimiento. Los reconocimiento de voz fueron llevados a cabo de la siguiente manera: todos los usuarios repitieron la misma frase diez veces a diferentes volúmenes de voz y sin importante ruido ambiental. En esta prueba se ha medido el volumen de la voz del usuario y la precisión del reconocedor. Los resultados son los siguientes:

- Volumen bajo - 69 dB: 0.66 valor de confianza (100 % tasa de acierto)
- Volumen medio - 77 dB: 0.72 valor de confianza (100 % tasa de acierto)
- Volumen alto - 83 dB: 8.80 valor de confina (100 % tasa de acierto)
- Volumen muy alto - 89 dB: 0.70 valor de confianza (100 % tasa de acierto)

Con estos resultados (ver Fig. 6.11) se puede concluir que aunque el volumen de la voz afecta al valor de confianza, las diferencias obtenidas no son muy elevadas, y sin embargo, la tasa de acierto sigue siendo la máxima en todos los casos. A la vista de esta prueba, se puede concluir que, cuando la pronunciación es clara y el volumen del audio es suficientemente alto como para que el humano lo pueda escuchar (sin llegar a saturar), la precisión de reconocimiento es óptima.

Después de varios años trabajando con herramientas de reconocimiento de voz, se ha podido observar que la mayoría de fallos en reconocimiento de voz se deben a problemas en la configuración del sonido, es decir, el sonido no llega claro al reconocedor de voz. Esto puede ser debido a que el micrófono esté mal conectado, los volúmenes de captura no sean los adecuados, haya acoples o interferencias, etc. La

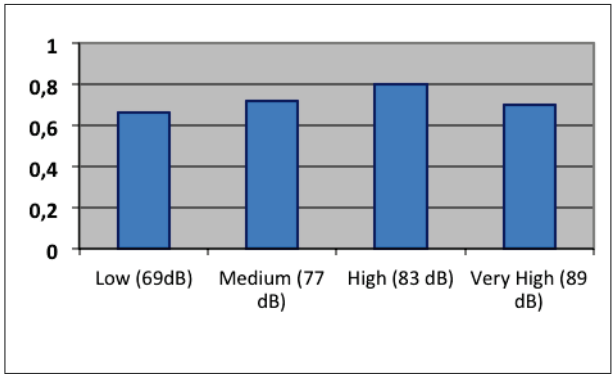


Figura 6.11: Precisión del reconocimiento con distintos volúmenes de voz

otra fuente principal de imprecisiones en el reconocimiento de voz, suele ser que las frases pronunciadas no se ajustan a la gramática establecida, bien porque no está fijada la gramática correcta para ese contexto comunicativo o bien porque la gramática no esté bien construida como para recoger el conjunto de frases posibles.

Diferentes entonaciones

Aquí se ha intentando determinar la relación entre la entonación de las frases y la precisión del reconocimiento. La misma frase es dicha por el mismo usuario con diferentes entonaciones: entonación enunciativa o "normal" y entonación interrogativa. Los resultados son los siguientes:

- Frases enunciativas: 0.74 valor de confianza y 97 % tasa de acierto.
- Frases interrogativas: 0.71 valor de confianza y 96 % tasa de acierto.

A la vista de estas pruebas, se puede decir que la entonación no es un factor decisivo que afecte al reconocimiento de voz. Aunque el tipo de oraciones que se usan con más frecuencia son las enunciativas con una emoción "normal" o de "tranquilidad", otro tipo de entonaciones y/o con otros estados emocionales (tristeza, nerviosismo, furia, etc.) también son correctamente reconocidas.

Diferentes géneros

Otro escenario importante es analizar cómo afecta el género del hablante a la precisión del reconocedor. Para ello, se han realizado pruebas de reconocimiento con hombres y mujeres de diferentes edades, usando la misma gramática, en un mismo entorno y con la misma configuración de micrófonos. Los resultados son:

- Mujeres: 0.70 valor de confianza y 99 % tasa de acierto.

- Hombres: 0.69 valor de confianza y 98 % tasa de acierto.

Con estos resultados se puede afirmar que el reconocedor es independiente del género del usuario que interactúa. Esto es lógico debido a que el modelo acústico del reconocedor ha sido entrenado por igual con hombres y mujeres.

Diferentes grupos de edad

En este caso se ha puesto a prueba el reconocedor de voz con grupos de gente de diferentes edades. Se ha dividido en dos grupos: niños de entre 5 y 12 años y adultos de 13 años en adelante. Los resultados obtenidos son:

- 5-12 años: 0.522 valor de confianza con una tasa de acierto del 93 %.
- 13-90 años: 0.722 valor de confianza con una tasa de acierto del 99 %

Se puede apreciar que en el grupo de edad comprendido entre 5 y 12 años la perdida de precisión de reconocimiento se hace notar. Esto es debido a que los niños se expresan usualmente con una voz menos formada, menos firme y en muchos casos incluso temblorosa o tímida. Sin embargo, la tasa de acierto todavía presenta valores bastante elevados. Este problema ha sido descrito en la literatura, por ejemplo en [Ishi et al., 2006a], y una solución adoptada es usar diferentes reconocedores para diferentes edades (entrenados de manera adecuada para cada grupo de edad).

En nuestro caso, dentro el sistema completo de diálogo, el tipo de interacción no queda únicamente restringido a voz, por lo que las personas con dificultad de interacción con el sistema de reconocimiento de voz pueden usar otros mecanismos alternativos de mayor tasa de acierto, como puede ser mediante etiquetas de radiofrecuencia dotadas de pictogramas.

6.9.2. Utilizando micrófono incorporado en el robot (omnidireccional)

En todos los escenarios descritos en la sección anterior se ha usado para las pruebas micrófonos auriculares inalámbricos unidireccionales. En esos casos, el usuario tenía que colocarse el micrófono cerca de la boca, lo que repercutía en una buena precisión de reconocimiento de voz pero a costa de perder un poco de naturalidad y comodidad en la interacción. Una interacción sin mecanismos adicionales externos al robot, como auriculares, es mucho mas natural y confortable [Breazeal et al., 2003].

En este nuevo escenario se prueba a realizar reconocimientos de voz con un micrófono omnidireccional (o no direccional) incorporado en el propio robot. Este tipo de micrófono es capaz de captar el audio del entorno en cualquier dirección. Normalmente son usados para captar el sonido en coros musicales, estadios deportivos, etc. Su

principal inconveniente, para nuestras necesidades, es que son mucho mas sensibles al ruido que los micrófonos unidireccionales, debido a que se han diseñado para propósitos diferentes a los nuestros. Sin embargo, su principal ventaja es que el usuario puede hablar al robot sin ningún tipo de artefacto adicional, consiguiendo una interacción muy similar a la que ocurre entre humanos, y por o tanto más natural.

Figura 6.12: Micrófono omnidireccional MP33865

Para probar la precisión del reconocimiento se han realizado pruebas a 1,2 y 3 metros de distancia al robot, sin importante ruido ambiental (menos de 50 dB). Los resultados obtenidos son:

- 1m: 0.42 valor de confianza (75 % tasa de acierto).
- 2m: 0.31 valor de confianza (72 % tasa de acierto).
- 3m: 0.25 valor de confianza (66 % tasa de acierto).

Como se puede ver, la precisión del reconocimiento decrece sustancialmente según aumenta la distancia de interacción. Esta perdida de precisión no sucede con el uso de micrófonos auriculares, puesto que la señal acústica se transmite inalámbricamente desde el emisor al receptor colocado en el cuerpo del robot. Con estos valores de precisión de reconocimiento y este tipo de micrófonos resulta difícil su uso en entornos reales, sin embargo en algún caso podría ser suficiente.

El segundo test realizado con estos micrófonos ha sido repetir la prueba anterior pero con importante ruido ambiental (65-75 dB), sin embargo en este caso ha sido imposible realizar ningún tipo de reconocimiento, dado que el motor de reconocimiento no es capaz de diferenciar entre ruido y voz. El problema radica en que el ruido y la voz llegan al detector de voz en volúmenes similares y la tecnología implementada de cancelación de ruido elimina ambas señales, siendo imposible detectar cuándo empieza y finaliza cada locución de voz. Además, este tipo de micrófonos carecen de mecanismos hardware de cancelación de ruido, porque no han sido diseñados para ese propósito.

Una solución para mejorar la interacción con estos micrófonos es reentrenar el modelo acústico del reconocedor de voz para esta nueva configuración. Recordad que el modelo acústico del reconocedor ha sido entrenado para aplicaciones telefónicas, por lo tanto usando micrófonos unidireccionales.

Concluyendo, este tipo de micrófonos eliminan la necesidad e incomodidad del uso de dispositivos externos al robot para la comunicación y proporcionan una manera de interacción mas natural entre el humano y el robot, sin embargo, en ambientes con importante ruido de fondo y/o en que la comunicación no es cercana entre ambas partes, son una pobre elección en términos de HRI.

6.9.3. Utilizando array de micrófonos

Un array de micrófono es cualquier número de micrófonos operando en tándem o paralelo (ver Fig. 6.13). Sus principales aplicaciones son extraer voz en ambientes ruidosos y localizar la fuente sonora (donde el usuario está situado espacialmente respecto al robot). Esas características son muy interesantes en HRI y han sido estudiadas recientemente en algunos trabajos [Kim & Choi, 2007] [Tamai et al., 2005] [Valin et al., 2004] [Yoshida et al., 2009]. Además, este moderno sistema de captación de audio está empezando a ser usado más y más en videojuegos [Chetty, 2009], ordenadores portátiles, teléfonos móviles, coches [Oh et al., 1992], etc. Casi todos ellos usan un array de micrófonos especialmente diseñado para dicho dispositivo, con entre 3 y 4 micrófonos unidireccionales, sobre los que se puede aplicar algoritmos de cancelación de ruido, cancelación de eco y localización sonora.

Estos sistemas son muy robustos frente al ruido. Combinan las ventajas del uso de micrófonos unidireccionales (poco sensibles al ruido de fondo) y las del uso de micrófonos omnidireccionales (se puede interactuar por voz con el robot sin la necesidad de auriculares, es decir, con libertad de movimiento).

Figura 6.13: Micrófono de array

Recientemente, están empezando a aparecer micrófonos de array comerciales y de propósito general en el mercado. Ellos están equipados con algoritmos de procesamiento de la señal que consiguen cancelación del ruido y localización sonora por hardware. Estos dispositivos son todavía relativamente caros y de un tamaño considerable. Para nuestras pruebas se ha adquirido uno de estos micrófonos de array, formado por 8 micrófonos y lo se ha integrado en el robot Maggie.

Nuevamente se ha repetido las pruebas de precisión de reconocimiento de voz para una distancia de entre 1 y 3 metros, sin importante ruido ambiental (menos de 50

dB). Los resultados obtenidos han sido:

- 1m: 0.62 valor de confianza (95 % tasa de acierto).
- 2m: 0.53 valor de confianza (83.78 % tasa de acierto).
- 3m: 0.37 valor de confianza (52.5 % tasa de acierto)

A los test anteriores se le añade importante ruido ambiental (65 dB aprox) y los resultados fueron:

- 1m: 0.47 valor de confianza (81,08 % tasa de acierto).
- 2m: 0.41 valor de confianza (80 % tasa de acierto).
- 3m: 0.31 valor de confianza (31,67 % tasa de acierto)

Se puede apreciar que la precisión disminuye según el usuario se ha ido alejando del robot, tanto en condiciones de silencio como en condiciones de ruido ambiental. En la siguiente sección se comparan estos resultados con los obtenidos previamente tanto por el uso de micrófono unidireccional y omnidireccional.

Pese a que para estas pruebas se ha usado un micrófono de array, formado a su vez por 8 micrófonos, también se ha desarrollado e integrado en el propio robot Maggie un sistema propio de 8 micrófonos colocados en forma de anillo en la base del mismo (ver Fig. 6.14).

Figura 6.14: Array de micrófonos en la base del robot

Este sistema propio de 8 micrófonos no está siendo actualmente usado para sistema de reconocimiento de voz, sino para tareas de localización de la fuente sonora, que se verán con detalle en otro de los capítulos de esta tesis doctoral. Los motivos de no usarse para tareas de reconocimiento de voz son dos. El primero de ellos, es que el

micrófono de array comercial entrega una única seña de audio (formada por la suma procesada adecuadamente de los 8 micrófonos), sin embargo, nuestro sistema de 8 micrófonos entrega 8 señales de audio diferentes, siendo difícil de tratar como una única señal. Este inconveniente podría ser salvado si en el propio sistema operativo (en este caso Ubuntu) se crea un único dispositivo virtual que englobe a estos 8 micrófonos. El otro inconveniente para su uso en tareas de reconocimiento de voz, es que estos micrófonos han sido colocados en la base del robot y la calidad de la señal recibida de la voz humana no es lo suficientemente buena como para realizar reconocimiento de voz y sí lo es, en cambio, para la tarea de localización de fuente sonora.

Resumen de las pruebas de precisión utilizando diferentes tipos de micrófonos

Para resumir el estudio realizado, se ha construido una gráfica que compara la precisión de los tres tipos de micrófono usados (ver Fig. 6.15). En esta figura, los tres sistemas de captura de audio son comparados en dos entornos: silencioso (menos de 50 dB) y ruidoso (65 dB aproximadamente). El sonido es capturado a una distancia igual a menor a 2 metros, en el caso de micrófono auricular, esta distancia es de apenas algún centímetro.

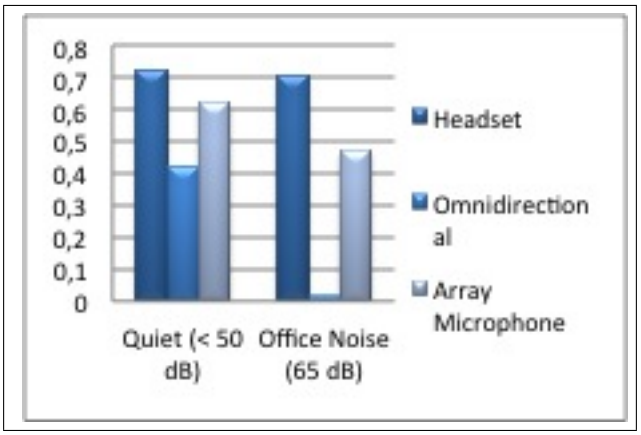


Figura 6.15: Precisión de reconocimiento usando distintos tipo de micrófonos

Con estos resultados, se puede afirmar que el uso de auriculares proporciona la manera mas precisa de realizar reconocimiento de voz. El siguiente sistema que proporciona mayor precisión es el de array; finalmente el peor de todos es el formado por un micrófono omnidireccional integrado en el propio robot. En los tres casos, se ha usado micrófonos de alta calidad (para el año 2010).

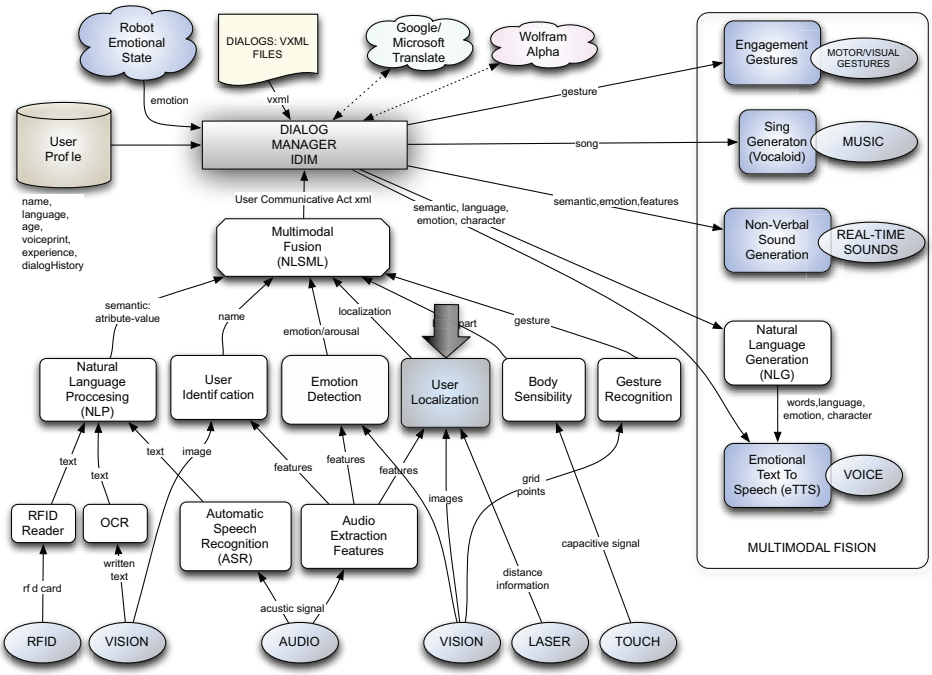
Dependiendo del tipo de interacción y el entorno, puede ser deseable usar un micrófono de array integrado en el propio robot o un micrófono auricular colocado en la boca del usuario. Si el entorno es muy ruidoso y/o es necesario altos valores de precisión en el reconocimiento de voz, el micrófono auricular es la mejor solución. Si el ambiente es mas silencioso, como pueda ser un hogar, y lo que interesa es una interacción lo mas natural posible, el array de micrófonos puede ser la mejor elección.

6.10. Resumen

En este capítulo se han descrito los pasos seguidos y los requisitos establecidos para dotar a nuestra arquitectura de control, y por ende, a nuestros robots, de la capacidad de entender el lenguaje natural hablado por cualquier humano y sin entrenamiento previo con el sistema. Se han analizado las diferentes alternativas de configuración hardware y sofware poniéndolas a prueba en entornos reales. En este sentido, se ha obtenido información sobre la precisión del reconocimiento en diferentes escenarios, usando los más modernos sistemas de adquisición de audio y analizando cómo influyen diversos parámetros en la precisión del reconocimiento: edad, sexo, entonación, volumen, idioma... Finalmente, se ha propuesto un nuevo modelo de clasificación de los resultados del reconocimiento como aceptados o rechazados, basándonos en un mecanismo de "segunda opinión", que utiliza dos reconocedores concurrentemente. Esta nueva aproximación tiene en cuenta las tasas de acierto de cada reconocedor pre-calculadas por cada intervalo de ruido, con ello se incrementa el numero de reconocimientos clasificados correctamente.

CAPÍTULO 7

Proxémica y sistema de localización multimodal del usuario



"Si me necesitas, dame un silbidito."— anónimo

7.1. Introducción

La supervivencia de los animales depende, en gran medida, de su efectividad percibiendo el entorno, en continuo cambio, en el que vive. Por lo tanto, la supervivencia depende en gran medida de su habilidad de actuar con la información sensorial que percibe, moviéndose hacia la comida o evitando posibles riesgos. Los robots autónomos móviles, especialmente aquellos diseñados para trabajar en ambientes hostiles, necesitan del diseño de capacidades análogas. En cambio, en el ámbito de los robots sociales, la percepción del entorno es usada para mejorar la interacción natural con el mismo, especialmente con los seres humanos, en tareas propias de los seres sociales como es el diálogo.

Los animales utilizan para detectar el origen de una fuente sonora el sentido del oído, sin embargo el número de "sensores" (oídos) empleados en esta tarea no es uniforme. Existen algunos invertebrados con sólo un oído, así como el rango de frecuencias audibles también varía. Las amantis religiosas están dotadas de un único oído, que usa como mecanismo de defensa para evitar la localización por ecos de frecuencias de los murciélagos [Yager & Hoy, 1986]. En ese sentido, los mamíferos y los invertebrados usan su sistema auditivo principalmente para tareas primarias, como evitar los peligros que le rodean, buscar alimento y aparearse; quedando excluidas funciones más avanzadas, propias de los seres humanos, como es la de entender el lenguaje natural o socializarse con el resto de sus semejantes.

En este capítulo, se presenta el desarrollo de un sistema de localización de usuarios basado en la fusión de información sensorial sonora y visual (como complemento a la sonora), integrado en el robot social Maggie.

Uno de los requisitos para lograr una interacción natural mediante diálogos entre humanos, o entre humanos y robots, es el de encontrarse en una situación espacial adecuada, es decir, dialogar a una distancia y orientación adecuada entre ambas partes para mantener proceso comunicativo satisfactorio. Los robots sociales suelen usar un sistema completo de diálogo multimodal que gestiona la interacción entre los usuarios y el robot durante el proceso comunicativo. Uno de los componentes que se han desarrollado, para este sistema de diálogo, es el relativo a la localización de usuarios respecto al robot. Este sistema proporciona información para determinar la situación espacial más adecuada del robot respecto al usuario. El determinar la situación más adecuada, requiere además de la localización del interlocutor, de un estudio proxémico en la interacción humano-robot (HRI).

En los últimos años, inspirándose en los estudios sobre el funcionamiento del sistema auditivo humano, se han trasladado al campo de la robótica móvil ([Hudspeth, 1983, Brown, 1874] y animal [Dooling & Popper, 2000, Ross & Smith, 1978]). En ese sentido, un nuevo campo de investigación a surgido, y se le conoce como Robot Audition [Nakadai et al., , Nakadai et al., 2002, Valin et al., a,

Murray et al., 2004, Andersson et al., 2004]. La mayoría de trabajos citados se centran en robots móviles y no en robots sociales. En los robots móviles el principal objetivo es el seguimiento de la fuente sonora (phonotaxis) y no el de situarse en el espacio, a una distancia adecuada para la interacción o diálogo (proxémica).

Los robots sociales, son artefactos diseñados para interactuar dentro de la sociedad, por lo que es importante que acaten sus reglas de comunicación, como son el respeto de los espacios. Además, es deseable que posean las características más avanzadas de su sistema auditivo, de una manera similar al humano. Los seres humanos tenemos dos oídos, con los que "sentir" el entorno sonoro en estéreo. Es en el cerebro donde, gracias a las diferencias en fase e intensidad del espectro de las señales acústicas percibidas por los oídos, se realizan las funciones de localización de fuente sonora.

El principal objetivo, dentro del área de investigación denominada *Robot Audition* aplicada a robots sociales, es el de mejorar la interacción del humano con el robot, durante el proceso de diálogo. Dicho diálogo se lleva a cabo gracias a un complejo sistema, formado por varios módulos independientes, pero que trabajan de manera coordinada para conseguir una interacción lo más natural posible entre el humano y el robot. Uno de estos componentes, en un sistema de diálogo moderno, es la localización de la fuente sonora, ya que puede ayudar considerablemente a situarse el robot espacialmente a una distancia adecuada para la interacción. Si bien, este componente es fundamental para dicha localización del robot en el espacio, no es el único factor o módulo que interviene para determinar dicha localización espacial. Para el robot social Maggie, se ha diseñado, desarrollado, implementado y probado en el sistema de diálogo que nos ocupa, en el que la localización de la fuente sonora desempeña un papel fundamental para la disposición espacial del robot.

7.2. Localización de la fuente sonora y de usuarios en robótica social

7.2.1. El problema de la localización sonora

En robótica, un sistema auditivo artificial puede ser usado para tres cosas fundamentales: 1) localización de las fuentes sonoras, 2) separar las fuentes sonoras en diferentes canales, 3) extracción de las características sonoras para realizar tareas como reconocimiento del lenguaje, detección de emociones o identificación del hablante. Este capítulo se focaliza en el primer punto, el resto se hace en otros capítulos.

Para lograr el objetivo de localizar las fuentes sonoras, según los trabajos [Wright & Fitzgerald, 2001] [Handzel & Krishnaprasad, 2002] existen varias aproximaciones al problema:

1. Una de ellas es estudiar las diferencias en amplitud que genera una fuente sonora entre los distintos micrófonos (u oídos) que reciben la señal. Este método, que es el que se sigue en este trabajo, se basa en comparar las diferencias de volumen entre los distintos micrófonos para determinar la diferencia angular con respecto a la fuente sonora. El micrófono más próximo a la fuente sonora, debería recibir un señal de mayor nivel de amplitud, que el resto de micrófonos, si bien, la precisión de este método se ve fuertemente afectada por los rebotes de la señal sonora contra los elementos del entorno, como paredes y muebles.

2. Un segundo método consiste en el análisis de las diferencias en fase que se producen entre las distintas señales recibidas por cada uno de los micrófonos relativas a la misma fuente sonora. Este último método tiene que ver, con que la misma señal generada por la fuente sonora será percibida instantes antes por el micrófono más cercano que el resto. La precisión de este método viene dado por el tamaño y la disposición del sistema de micrófonos, si los micrófonos están situados demasiado cerca, todos ellos recibirán prácticamente la misma señal [Brandstein & Ward, 2001, Benesty et al., 2008]. La primera de las técnicas trabaja en el dominio del tiempo, mientras que la segunda trabaja en el dominio de la frecuencia. Ambas técnicas se pueden combinar para lograr un sistema de mayor precisión.

3. Existe una tercera técnica, no tan comentada bibliográficamente debido a su complejidad, y que sólo necesita de un micrófono para su desempeño [Saxena & Ng, 2009]. Consiste en el estudio de las diferencias en el espectro, producidas por la misma fuente sonora emitiendo el mismo sonido, desde distintas posiciones relativas al micrófono. Los seres humanos, que poseemos únicamente dos oídos, somos capaces de diferenciar si un sonido viene de delante o de detrás, incluso si se produce a la misma distancia y ángulo tanto por delante como por detrás. Esto es posible precisamente gracias a esta tercera propiedad que se está analizando, que posibilita a los humanos únicamente con un solo oído realizar localización sonora. Sin embargo, para un sistema de *audición artificial* basado en un sólo micrófono, resulta un desafío, ya que es necesario de cierto conocimiento a priori del sonido, para poderlo comparar con el sonido recibido previamente desde otra posición. Ciertos sonidos, que se escuchan en nuestro entorno a diario, tienen cierta estructura, que en el caso de los humanos somos capaces de reconocer. En los humanos este conocimiento del sonido se va ganando durante años de experiencia [Morrongiello, 1989].

El uso combinado de la primera y segunda técnica, se está haciendo habitual en el campo de la robótica [Jwu-Sheng et al., 2009, Andersson et al., 2004, Valin et al., a, Nakadai et al., , Nakadai et al., 2002, Briere et al., 2008]. Para conseguir resultados

satisfactorios aplicando el algoritmo que trabaja con diferencias en fase de la señal (segunda técnica), se requiere de ciertas especificaciones hardware que limitan su uso: una disposición de los micrófonos, en la que cada uno de ellos quede suficientemente alejado como para poder captar diferencias en fase entre las señales percibidas, lo que implica, robots de cierto tamaño en los que poder situar los micrófonos correctamente. Se pueden encontrar sistemas de array de micrófonos comerciales [1] de no menos de 11 pulgadas, lo que puede resultar incomodo y poco estético. Otro problema añadido es la necesidad de una tarjeta que lea físicamente todos los micrófonos simultáneamente.

En el campo de la robótica, se han desarrollado dos paquetes de software básicos que cubren los tres apartados descritos previamente. Si bien, para este estudio sólo nos interesa su aportación al sistema de localización de fuente sonora. Estos dos frameworks son ManyEars [Peto, 1980] y HARK [Nakadai et al., 2008b, Takahashi et al., 2010] que implementan algoritmos de filtros de partículas [Valin et al., 2007] para localización de forma robusta, que se basan en los dos métodos comentados previamente (diferencias en amplitud y fase). Según los autores, logran una precisión de la localización sonora entorno a los 3º de margen de error.

El principal problema del uso de estos entornos es que son muy complejos y la necesidad de una tarjeta específica de lectura simultánea de todos los micrófonos. Para nuestras necesidades, dentro del campo de la robótica social, y gracias a la fusión sensorial realizada, nos es totalmente suficiente con un sistema más sencillo, basado en diferencias en amplitud y sin la necesidad de una tarjeta específica de adquisición.

7.2.2. Fonotaxis *vs* proxémica

Fonotaxis, se puede definir como "el movimiento de un organismo en relación a la fuente sonora. Por ejemplo, las hembras de ciertas especies animales, se sienten atraídas por el canto de cortejo de una pareja potencial (es decir, positiva phonotaxis), o algunos animales pueden huir al oír el sonido de un posible depredador (es decir, negativa phonotaxis)"[2]. La phonotaxis permite discriminar donde está situado el foco de atención al que dirigirse, basándose únicamente en el sentido del oído. En ese sentido, es muy diferente el papel desempeñado por el sistema de audición en un robot social, que el papel desempeñado en la especie animal o en ciertos robots móviles, cuyo principal objetivo es cubrir sus necesidades primarias, entre ellas la de supervivencia.

Por otro lado, la **proxémica** se define como la parte de la semiótica[3] dedicada al estudio de organización espacial durante la comunicación lingüística [Hall, 1966]. Más

[1] http://www.acousticmagic.com/

[2] A dictionary of Biology: http://www.encyclopedia.com/doc/1O6-phonotaxis.html

[3] es el estudio de los signos y señales usados en la comunicación

concretamente, la proxémica estudia las relaciones de proximidad, de alejamiento, etc. entre las personas y los objetos durante la interacción, las posturas adoptadas y la existencia o ausencia de contacto físico. Asimismo, pretende estudiar el significado que se desprende de dichos comportamientos.

La competencia proxémica permite a las personas crear un marco de interacción acorde con unas coordenadas espacio-temporales que expresan determinados significados y que, en ocasiones, obedecen a un complejo sistema de restricciones y normas sociales que tienen que ver con el sexo, la edad, la procedencia social, la cultura, etc. Por otro lado, a veces la distribución del espacio está establecida de antemano, por ejemplo, en la sala de un juicio o en una ceremonia religiosa.

La finalidad del sistema de localización sonora en un robot social, como Maggie, es complementar al sistema de diálogo multimodal, en una tarea más compleja que tiene que ver con la proxémica y es adecuada para una interacción natural. El gestor de diálogo puede gestionar la información de localización sonora con distintos fines, uno de ellos puede ser realizar "phonotaxis", pero cualquier otro es posible, como por ejemplo alejarse del usuario si este actúa con agresividad, o mejorar el enganche (*engagement*) en el proceso comunicativo, etc.

7.2.3. Análisis proxémico en la interacción entre humanos

El origen de la proxémica está relacionado con los estudios que los etólogos [4] habían realizado acerca de la importancia de la distribución espacial en las interacciones entre animales. En los años setenta, un grupo de investigadores, entre ellos el antropólogo Edward T. Hall [Hall, 1966], aplicaron el modelo que etólogos, como Huxley o Lorenz, habían diseñado para el mundo animal al estudio de la comunicación en las sociedades humanas, e introdujo el concepto de proxémica. Hall identificó varios tipos de espacio, entre ellos el denominado espacio personal. Este espacio no es otro que el creado por los participantes de una interacción y que varía en función del tipo de encuentro, la relación entre los interlocutores, sus personalidades y otros factores. Diseñó un modelo en el que clasifica el espacio personal en cuatro subcategorías:

- **Espacio íntimo**, que va desde el contacto físico hasta aproximadamente 45 cm. Esta distancia podría subdividirse en dos intervalos distintos: entre 0 y 15 cm, distancia que presupone el contacto físico y que tendría lugar en situaciones comunicativas de máxima intimidad (por ejemplo, durante el mantenimiento de relaciones afectivas); y entre 15 y 45 cm, que se corresponde con una distancia menos íntima, pero si dentro de un marco de privacidad.

- **Espacio causal-personal**, que se extiende desde 45 a 120 cm. Es la distancia

[4] investigadores del comportamiento comparado entre el hombre y los animales

habitual en las relaciones interpersonales y permite el contacto físico con la otra persona.

- **Espacio social-consultivo**, que abarca desde los 120 cm hasta los 364 cm. y aparece en situaciones donde se intercambian cuestiones no personales.

- **Espacio público**, que va desde esta última hasta el límite de lo visible o lo audible. A esta distancia los participantes tienen que amplificar recursos como la voz para posibilitar la comunicación, esto sucede por ejemplo durante una conferencia.

El propio E. T. Hall señala que este modelo está basado en sus observaciones de una muestra particular de adultos y por lo tanto, no es generalizable a todas las sociedades. Es evidente, que existen normas diferentes en cada cultura para el lugar y distancia que se deben mantener en determinadas situaciones y que transmiten información sobre la relación social entre los participantes. Existe una distancia adecuada para cada situación de acuerdo a unas reglas establecidas por la comunidad que los participantes conocen, o deben aprender, para moverse con éxito en las relaciones interpersonales y evitar conflictos o interpretaciones erróneas.

Otros importantes trabajos de la época [Argyle, 1988, Argyle, M.; Dean, 1965], analizan y experimentan con distintos factores que influyen en la proxémica, como son la mirada, personalidad, familiaridad, número de personas interactuando, normas culturales, etc. En ellos se explican la necesidad de contacto visual para la interacción. Contactos visuales prolongados de más de 10 segundos aumentan la ansiedad, sin embargo la ausencia de contacto visual directo hace que la gente no se sienta integrada completamente en la conversación.

En un trabajo muy preciso [Lambert, , Lambert, 2004], Lambert estableció mediadas de distancia comunicativa en función del tipo de situación afectiva que se presentaba (ver Fig. 7.1). Si bien, dejaba de lado muchos otros aspectos que tienen también que ver en la interacción humana y que influye directamente a dicha situación espacial.

Los seres humanos, al igual que los demás animales, manejan el espacio y emplean las distancias como una manera de satisfacer necesidades vitales y de relación con los demás; sin embargo, los estudios proxémicos han podido establecer que la percepción que tenemos del espacio personal y del espacio social resulta culturalmente determinada [Hayduk, 1978]. Biología y cultura, pues y como en tantas otras cosas, se combinan en la utilización que hacemos del territorio.

Range	Situation	Personal Space Zone
0 to 0.15m	Lover or close friend touching	Intimate Zone
0.15m to 0.45m	Lover or close friend only	Close Intimate Zone
0.45m to 1.2m	Conversation between friends	Personal Zone
1.2m to 3.6m	Conversation to non-friends	Social Zone
3.6m +	Public speech making	Public Zone

Figura 7.1: Espacios personales en la interacción humano-humano según Lambert

7.2.4. Análisis proxémico en interacción humano-máquina y humano-robot

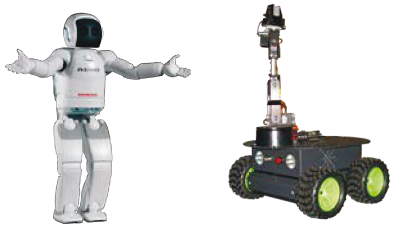
Se han realizado estudios proxémicos en interacción Humano-Máquina [Nass, C; Reeves, 1996, Nass et al., 1994]. En estos trabajos, los usuarios interactúan con computadoras o agentes virtuales, por lo que los resultados no pueden ser extrapolados a un estudio de la interacción, en los cuales los usuarios interactúan con ordenadores, agentes virtuales, etc. Si bien estos estudios, en los que el agente virtual carece de cuerpo físico (embodyment), difícilmente pueden ser extrapolados al campo de la interacción humano-robot, ya que el modelo de interacción entre el humano y el robot no suele estar centrado en el uso de pantalla y ratón.

El uso de la localización sonora, en el ámbito de un robot social y su uso dentro del campo de la proxémica constituye una tarea mucho más rica y ambiciosa que las meras capacidades de "phonotaxis" implementadas en sistemas de localización sonora en robots móviles. Una correcta implementación de un sistema de diálogo proxémico, permite mantener interacciones a la distancia adecuada al interlocutor o los interlocutores, variando en orientación y distancia en función de múltiples características como son: identidad del interlocutor, número de interlocutores, distancia y ángulo previo a la interacción, estado afectivo del propio robot y de los interlocutores ([Takayama & Pantofaru, 2009]). Todos estos aspectos, únicamente pueden ser tratados, en interacciones reales, si se dispone de un completo sistema de diálogo dotado de suficientes módulos capaces de realizar dichas tareas.

En la literatura, se puede encontrar varios estudios sobre proxémica y HRI que analizan alguno de esos factores. Por ejemplo, Breazeal[C. Breazeal, 2000] encontró que el usuario respondía a robots zoomórfícos de formas no verbales (con sonidos) y respetando el espacio interpersonal del robot. Por otro lado, Hüttenrauch [Hüttenrauch, 2006] observó que en pruebas de HRI la mayoría de los participantes guardaban distancias de interacción correspondientes al espacio 2 de Hall (0.45 to 1.20 m.). En otro estudio [Walters, 2009] encontró que los participantes generalmente permitían que el

robot se aproximase a ellos en las interacciones físicas mucho más que en interacciones meramente verbales.

En ([Kheng Lee et al., 2007]) se realiza un interesante estudio sobre la apariencia externa del robot y la proxémica. Relacionó la proxémica y la experiencia de uso ganada (interacción a corto o largo plazo), viviendo con el robot por un periodo de tiempo superior a cinco semanas. Durante su primer encuentro, los participantes del estudio exhibieron una fuerte tendencia a permitir que el robot, de apariencia mecatrónica, se aproximase más cerca que los robots con apariencia humana (ver Fig. 7.2.4). Por lo tanto, demostró que la distancia de interacción depende en gran medida de la apariencia del robot, en este caso humanoide o mecatrónica. Los usuarios generalmente preferían estar más cerca de los robot mecatrónicos que de los humanoides, aunque esta tendencia se iba diluyendo según los participantes se acostumbraban al uso del robot.



(a) Humanoide (b) Mecatrónico

Figura 7.2: Diferencia de aspecto externo entre un robot humanoide y uno mecatrónico

En [Takayama & Pantofaru, 2009] varios factores fueron analizados. Descubrió que la mirada de un robot hacia el usuario también influye en los comportamientos proxémicos. Además, se encontró que las mujeres mantenían mayores distancias de interacción que los hombres, cuando el robot las miraba. Factores como aptitudes, experiencias personales y personalidad también tienen su influencia directa sobre el comportamiento proxémico. Las personas con comportamiento más agradable tendían a colocarse más cerca del robot, a diferencia de las personas con actitudes mas negativas o neuróticas hacia los robots en general tendían a situarse mas alejados de los robots. Siguiendo esta misma línea, en [Mumm & Mutlu, 2011] se mostraba como los participantes que les gustaba el robot, mantenían la distancia de interacción aunque el robot les mirase directamente, en cambio, los participantes que no les gustaba el robot, si este les miraba, mantenían mayores distancias de interacción que sino se sentían intimidados por su mirada.

Debido a que algunos comportamientos no-verbales pueden ser considerados elementos opcionales a añadir en un sistema de diálogo multimodal, existen muy pocos trabajos en el que realmente la capacidad proxémica de decidir la posición adecuada para el proceso comunicativo está incluido dentro del propio sistema de diálogo [Henkel, 2012]. Los trabajos previamente citados se centran en el uso de la proxémica fuera del sistema de diálogo, pero para lograr una interacción humano-robot natural, es esencial que esta sea integrada dentro del sistema de interacción. Siendo este último, el que teniendo en cuenta toda la información que dispone, tipo de usuario, perfil, etc., pueda relacionarla para decidir la disposición mas adecuada.

7.3. Factores en análisis proxémico entre el usuario y el robot social Maggie

Como se ha visto en la sección 8.2, Hall realizo estudios de la proxémica entre seres humanos, fundamentalmente entre personas adultas. Partiendo de esa base y de los cuatro espacios que identificó, se ha tratado de llevar ese mismo estudio al tipo de interacciones que se han producido entre el robot social Maggie y chicos de 8 a 17 años, que provenían de varios colegios.

Los experimentos fueron realizados en grupos de 15 alumnos, interactuando con Maggie en sesiones de 15 minutos. Un total de 60 alumnos han participado en este trabajo. Interactuaron en diferentes situaciones: en grupos, uno a uno y con/sin ayuda de un experto. Además, ninguno de ellos tenía experiencia previa con el sistema ni con el robot. Para probar la interacción con adultos, se realizaron test con 5 miembros del grupo de desarrollo del RoboticsLab, con edades comprendidas entre 25 y 30 años. En este caso, todos ellos tenían experiencia previa con el uso del robot Maggie.

El objetivo de estos estudios fue identificar como los espacios personales, descritos por Hall se relacionan con los diferentes tipos de usuarios y situaciones comunicativas. Esta información es esencial para que el sistema de dialogo (en concreto el gestor del diálogo) tome la decisión de colocar el robot en una situación adecuada respecto al usuario. En otras palabras, para adaptar la distancia de interacción al tipo de usuario/s. Los aspectos tenidos en cuenta son: experiencia de uso (familiaridad con el sistema mediante el uso de perfiles de usuario), la edad, la personalidad, el género, el aspecto externo del robot y el número de usuarios.

Todas las interacciones fueron grabadas en vídeo para analizar los diferentes factores que podían influir en la proxémica en la interacción entre los usuario y el robot. En estas interacciones los usuarios podían interactuar libremente usando todos los diálogos y posibilidades de interacción ofrecidas por Maggie, por lo tanto, sin restricciones en su uso. En el caso de niños, cada interacción fue absolutamente diferente, y normalmente intentaron jugar con el robot (una descripción extensa de los juegos de

Maggie puede encontrarse en [Gonzalez-Pacheco et al., 2011]).

7.3.1. Experiencia de uso

Se ha considerado dos tipos de interacciones con el robot social: con y sin perfiles de usuario. En el primero de ellos, el usuario interactúa con el robot sin tener un perfil creado, típicamente porque es su primera interacción con el robot (quizás durante una demostración o show). Por lo tanto, el comportamiento del robot se focaliza en llamar la atención del usuario. Si el usuario quisiera interactuar con Maggie varias veces, sería conveniente que se registre en el sistema, para que el robot lo conozca y pueda adaptar su comportamiento. En el perfil de usuario se guardan varias características: nombre, edad, género, experiencia de uso, huellas de voz y lenguaje.

Gracias al uso de perfiles, el diálogo puede adaptarse al usuario, y una de estas adaptaciones tiene que ver con la distancia de interacción. Antes de cargar el perfil, es necesario identificar al usuario. En nuestro sistema de diálogo, dicha identificación se realiza mediante voz (hasta ahora la visión no es usada para complementar dicha identificación). Para lograrlo, durante la fase de registro, el sistema aprende las características de la voz del usuario y las almacena en un fichero, al que llamamos "huellas de voz". Estas huellas de voz son necesarias para identificar el usuario cuando saluda al robot.

En [Takayama & Pantofaru, 2009], la relación entre la experiencia de uso con el sistema y la proxémica es analizada. De hecho, se ha observado que los usuarios con perfil, y por lo tanto con experiencia con el sistema, es común mantener una distancia de interacción que se corresponde con el espacio 3 de Hall (120 a 364 cm). Sin embargo, durante una interacción espontanea, cuando el robot está haciendo un show o el usuario no está experimentado, por lo tanto sin perfil de usuario, la distancia de interacción se relaciona con el espacio 4 de Hall (espacio público).

Finalmente, en interacciones entre humanos, el rango de familiaridad parece estar relacionado con una menor distancia de interacción [Hall, 1966]. Sin embargo, considerando la experiencia ganada con el robot como una media de familiaridad con el robot, se ha observado que la distancia de interacción crece según se incrementa la experiencia. Esto parece ser debida a la propia naturaleza de interacción con Maggie.

En la comunicación con Maggie, el reconocimiento de la voz es usado normalmente mediante micrófonos auriculares inalámbricos colocados en el usuario, mientras que la localización sonora se realiza por los 8 micrófonos colocados en la base del robot. Por lo tanto, según aumenta la experiencia con el robot, los usuario se dan cuenta que pueden alejarse del robot sin que por ello disminuya la precisión del reconocimiento de voz. Sin embargo, un usuario novel, tiende a colocarse cerca del robot para que el robot lo escuche mejor, situación que se ve acrecentada si se producen varios fallos de reconocimiento de voz.

Estas conclusiones han sido observadas durante sesiones donde los chicos y los miembros del equipo interactuaban individualmente con el robot.

7.3.2. La edad

Analizando los datos recogidos, se ha observado que la edad también influye en la distancia de interacción. De media, la distancia de interacción con niños de entre 8 y 10 años, se sitúa por encima de los 2 metros (espacio personal 3), ver Fig. 7.3. Sin embargo, para niños mayores de 10 años y adultos, la distancia de interacción disminuye a 1m, ver Fig. 7.4. Nosotros pensamos que los niños con menos de 10 años se sienten más intimidados por el robot que niños mayores. Los chicos mayores sienten mayor grado de curiosidad hacia el robot y se sienten menos intimidados, por lo que se aproximan en mayor medida al robot. En el caso de las personas adultas que ya han probado la interacción con el robot, se sienten confortables con la interacción.

Figura 7.3: Interacciones con niños de entre 8 y 10 años

Figura 7.4: Interacciones con personas mayores de 10 años

7.3.3. La personalidad

La personalidad es un factor que influye en la proxémica, según queda reflejado en trabajos como [Williams, 1971]. En los experimentos realizados se ha observado que en interacciones grupales (varios niños interactuando con el robot), son los más extrovertidos los que tienden a situarse más próximo al robot, mientras que los más tímidos, por su propia naturaleza son mas propensos a mantener una distancia de seguridad con el robot, buscando siempre el cobijo del tutor, sobre el que se sienten más seguros. Son además los alumnos mas extrovertidos los que más ganas tienen de interactuar con el robot, y tratando de llamar su atención y ganar el foco de atención del robot sobre el resto de niños, tienden a situarse más cerca del mismo. Medir cuantitativamente la personalidad de cada interlocutor es ciertamente una tarea difícil y poco precisa, pero parece claro que en interacción humano-robot el grado de timidez está bastante relacionado con la distancia de interacción y la duración del diálogo.

7.3.4. El género

Otro factor que puede influir en la proxémica, es el género del usuario. Estudios previos parecen reflejar que las mujeres prefieren situarse frente al robot y los hombres prefieren situarse lateralmente al mismo [Fisher & Byrne, 1975]. En cambio, en nuestros estudios, no se ha podido corroborar esta afirmación, ya que no se ha apreciado ciertamente ninguna diferencia entre la manera de actuar de los niños y las niñas.

7.3.5. El aspecto del robot

Hasta ahora los factores que se han analizado, tienen que ver con factores propios a cada interlocutor, sin embargo, existen otros factores, no de menor importancia, que también influyen en la proxémica. Estos otros factores tienen que ver con la propia naturaleza del robot, esto es: forma del robot, colores, altura, aspecto, volumen de la voz, tipo de comunicación (verbal, sonora, gestual), peso, velocidad de traslación, etc. Estos aspectos no se han analizado, fundamentalmente porque únicamente se ha experimentado con el robot Maggie.

7.3.6. El número de usuarios

Aunque el sistema de diálogo está diseñado para interactuar con los usuarios de uno a uno, es decir, no es posible cargar varios perfiles de usuario simultáneamente, cualquiera de ellos puede hablar al robot y llevar a cabo algunos procesos comunicativos en modo cooperativo. Por esta razón es interesante estudiar la interacción en grupos.

Se ha observado que, durante la interacción con mas de un chico, los chicos tienden a situarse muy cerca del robot, tratando de llamar su atención sobre el resto de los miembros del grupo. De hecho, se ha observado que el mismo niño que empieza interactuando él sólo con el robot a una distancia lejana, se aproxima al robot cuando sus compañeros de clase se unen a la interacción 7.5.

Figura 7.5: Interacción grupal

Por otro lado, en tareas que requieren coordinación, por ejemplo cuando el robot comienza a cantar y bailar, se ha observado que algunos niños, de manera espontánea, han tendido a alinearse con el robot e imitar sus pasos de baile. Por lo tanto, su disposición espacial respecto al robot cambia (ver 7.6).

Figura 7.6: Niños imitando al robot bailando

7.3.7. Conclusión: reglas proxémicas observadas en la interacción usuario-Maggie

En esta sección, se presenta el conjunto de reglas extraídas fruto del estudio proxémico realizado en la interacción de los usuarios con el robot Maggie. Estas reglas son aplicadas en nuestro propio sistema de diálogo. No obstante, todos los factores analizados no se han podido tener en cuenta a la hora de determinar las reglas con

las que el sistema de diálogo debe decidir su disposición espacial. Por ejemplo, la personalidad del usuario no se ha podido traducir en ninguna regla concreta, puesto que el sistema de diálogo actualmente no tiene forma de determinarla. Además el genero tampoco se ha tenido en cuenta puesto que no se ha encontrado variaciones significativas entre la distancia de interacción de hombres y mujeres. Tampoco se ha tenido en cuenta el número de usuarios interactuando simultáneamente con el robot, debido a que el sistema de diálogo únicamente es capaz de cargar un perfil de usuario en cada interacción. Recordar que el perfil de usuario contiene la información concreta de cada usuario y permite la personalización/adaptación del diálogo (incluyendo la adaptación proxémica).

En la Fig. 7.7 se muestra como las reglas proxémicas son aplicadas durante la HRI. Como se puede observar, cuando el usuario saluda al robot, dos situaciones pueden presentarse: que el usuario sea identificado por el sistema (el robot tiene su perfil de usuario),o no (el sistema no conoce a dicho usuario y no tiene su perfil). En el primero de los casos, el robot carga el perfil del usuario correspondiente e incrementa el valor de la experiencia de uso con el sistema en uno. Por otro lado, si el usuario no es reconocido/identificado por el sistema, el robot le preguntará si quiere registrarse en el sistema. En el caso de que el usuario no quiera registrarse, el robot mantendrá una distancia de interacción de, al menos, 3.6 m (espacio personal 4). Por contra, si el usuario sí quiere registrarse en el sistema, un nuevo perfil de usuario es creado, y el robot mantiene una distancia inicial comprendida entre 120 y 364 cm (espacio personal 3). Desde entonces y dependiendo de la edad y el nivel de experiencia, la distancia varia (siempre dentro del espacio 3). Si el usuario tiene una edad de entre 8 y 10 años, la distancia que se mantiene es aprox. 250 cm; en otro caso, dependiendo de su nivel de experiencia (medido como número de interacciones con el sistema), las distancias varían entre 120 y 225 cm.

7.4. Descripción del sistema

Para integrar en el robot social Maggie, y para ser precisos, dentro de su sistema de diálogo, de la capacidad de localizar el o los usuarios en el entorno del robot, es necesario hacer una descripción software y hardware del problema.

7.4.1. Sistema *hardware*: sensores usados

Un sistema artificial de localización sonora, con únicamente dos sensores de sonido (micrófonos), es ciertamente impreciso. Es difícil diferenciar si el sonido es frontal o trasero, así como conseguir altos niveles de precisión. Sin embargo un robot, no está limitado a usar dos micrófonos; en ese sentido se ha decidido usar ocho micrófonos

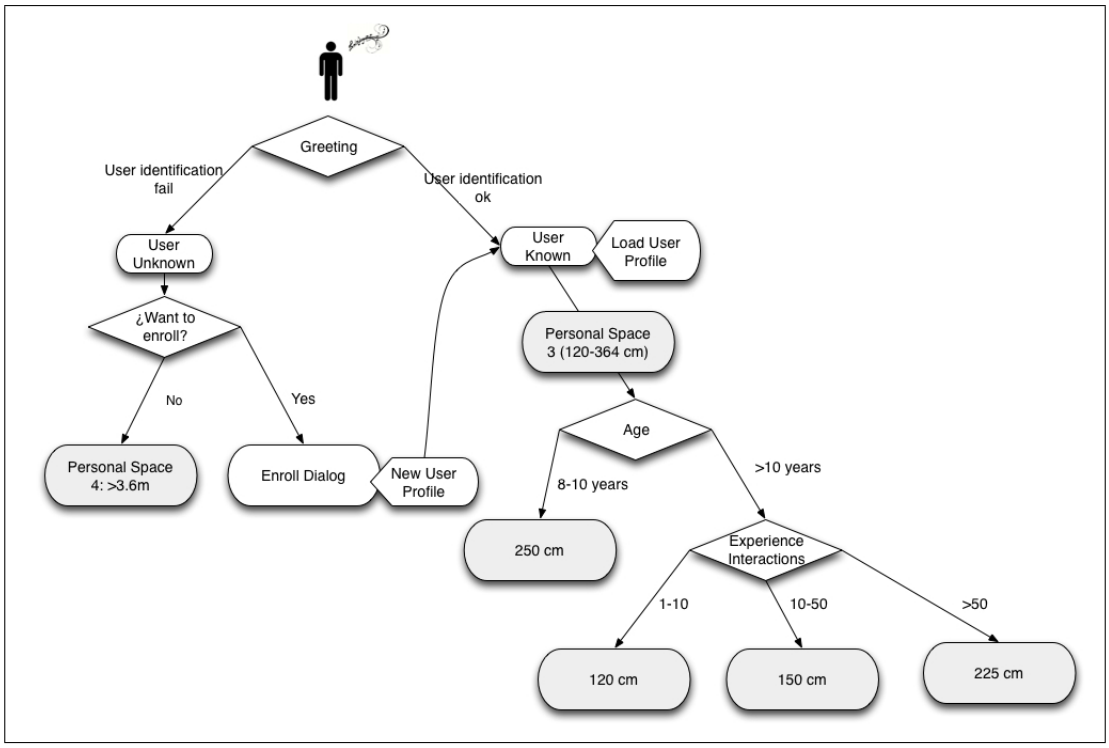


Figura 7.7: Reglas proxémicas aprendidas en interacciones reales entre usuarios y el robot social Maggie

colocados concéntricamente sobre la base del robot Maggie, esto repercute en una mayor robustez frente a los efectos del ruido.

En nuestro robot personal se ha integrado 8 micrófonos direccionales que se conectan por USB al ordenador incorporado en el cuerpo del robot. Estos micrófonos se colocan en la base del robot en una perfecta circunferencia de 40 cm de radio y a 21 cm del suelo. Esta distribución se puede ver en las figuras 10.1 y 7.9.

Figura 7.8: Micrófonos en el robot Maggie

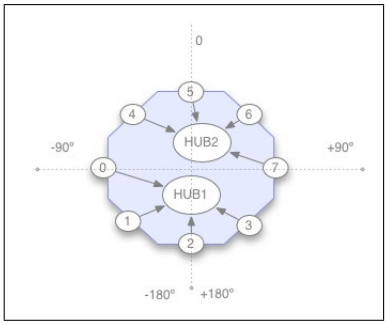


Figura 7.9: Esquema de situación de los micrófonos en Maggie

Para la extracción de características sonoras, necesarias para el reconocimiento de voz, la detección de emociones y la identificación de usuarios, se usa otro micrófono adicional, un micrófono inalámbrico auricular direccional, mucho menos expuesto al ruido de ambiente [Alonso-Martin & Salichs, 2011].

La localización de los micrófonos en la parte baja del robot, se ha decidido por dos factores: se alejan de las propias fuentes de sonido del robot (los altavoces) que están colocados en el cuello del mismo, y por ser la parte del robot con una perfecta simetría circular para favorecer los algoritmos de cálculo de localización de la fuente sonora.

Notar que situar los micrófonos en la propia estructura del robot facilita la tarea de localización sonora, puesto que el propio cuerpo actúa como barrera para las ondas sonoras sin visión directa a la fuente, De esta manera los micrófonos mas alejados de

la fuente sonora reciben una señal de mas baja intensidad que los situados en visión directa a dicha fuente. Si bien esta disposición y el hardware usado, nos ha generado dos inconvenientes con los que se ha tenido que lidiar: la lectura simultanea de los micrófonos, así como el acople de los propios sonidos generados por el robot durante el diálogo con el interlocutor. Estos problemas se analizaran a continuación.

En nuestro caso, se ha querido evitar dicho coste y se han usado micrófonos USB convencionales conectados a varios Hub USB, que a su vez se conectan al ordenador del robot. Los dispositivos de audio asociados a dichos micrófonos deben cargarse siempre en el mismo orden, para evitar su desordenación lógica en el sistema operativo que los lee.

Como se ha comentado, el sistema de localización de usuarios, no sólo usa el sonido como fuente de entrada de información, se apoya en información visual o de distancias obtenidas para disminuir el error cometido por el sistema auditivo. Para ello es necesario de un telémetro láser infrarrojo que nos proporciona información de distancias, así como el sistema de visión Kinect, comentado en la sección que describe el entorno de trabajo.

7.4.2. Sistema *software*

Nuestra arquitectura de control corre sobre un sistema operativo Linux, concretamente un Ubuntu 11.103, y una arquitectura de sonido ALSA. En el cual es posible ordenar los dispositivos de audio en un orden fijo, mediante el fichero de configuración /etc/modprobe.d/alsa-base.conf.

Primeramente ha sido necesario una fase previa de calibración. Aún siendo todos los micrófonos el mismo modelo, cada uno de ellos tiene distinto nivel de captura (ganancia intrínseca a cada micrófono), por lo que es necesario fijar, mediante prueba y error, un volumen de captura de audio uniforme en el sistema operativo para todos ellos.

La lectura se realiza secuencialmente sobre cada uno de los micrófonos, leyendo en cada interacción una pequeña cantidad de tramas, 256 es un buen valor, para que la lectura bloqueante por cada micrófono sea lo mas pequeña posible. Por cada interacción de lectura, se leen 256 tramas de audio de cada micrófono, siendo la interacción tan rápida que casi la lectura en un instante de tiempo se acerca bastante a una lectura de datos simultanea (menor a 30ms).

Sobre estas tramas, usando las funciones propias de ALSA, se calcula el nivel de intensidad sonora que se alcanza por cada uno de los micrófonos leídos. Este proceso se repite durante la lectura de un número dado de iteraciones, en nuestro caso 5, y se calcula un valor medio de intensidad de señal leído por cada uno de los micrófonos. Si en lugar de usar 5 iteraciones se usa una cantidad mayor, el sistema será menos "sensible" y por lo tanto también menos "reactivo" a cambios sonoros en el entorno,

ya que el calculo promedio llevará mas tiempo que para un numero de iteraciones inferior.

Una vez que se ha calculado un valor de intensidad medio para cada micrófono durante una serie de iteraciones fijas, se comprueba cual de los micrófonos es el que registra un mayor nivel de intensidad. Si ese nivel de intensidad supera un determinado umbral, que se ha fijado previamente para discriminar sonidos de fondo, de voz humana o sonidos relevantes, y además el robot no se encuentra hablando en ese momento, puesto que la propia voz del robot podría ser la fuente sonora, entonces se determina que la fuente sonora se encuentra en el ángulo que corresponde al micrófono que ha recibido una mayor intensidad de señal acústica. Ver el algoritmo 7.1.

Algoritmo 7.1 Sound source localization algorithm

Require: $numMicrophones$ =8, $numSamples$ =256, $numIterations$ =5, $voiceThreshold$ =1100

```
1:  int frames[numMicrophones][numSamples]
2:  int accumulatedVolume[numMicrophones]
    {The volume is computed or each microphone in several iterations}
3:  for numIter ← 0 to numIterations do
4:     readAudioSamplesAllMicrophones(frames)
5:     for numMicro ← 0 to numMicrophones do
6:        for numSample ← 0 to numSamples do
7:           accumulatedVolume[numMicro] += frame[numMicro][numMuestra]
8:        end for
9:     end for
       {Look for the microphone with more accumulated volume}
10:    int microphoneWin = getMaximo(accumulatedVolume)
       {If robot is not speaking and accumulated volume of microphoneWin is upper
       the voiceThreshold}
11:    if (accumulatedVolume[microphoneWin] ≥ voiceThreshold) AND robotIs-
       Quiet() then
12:       int angleSoundSource = (360/numMicrophones)*microphoneWin
13:       emit(angleSoundSource)
14:    end if
15: end for
```

Después de que el sistema de localización sonora determina la orientación del usuario respecto al robot, el sistema de localización basado en medidas de distancia del láser empieza a trabajar. Este sistema permite medir las distancias de interacción respecto al usuario, proporcionando mucho mas precisión que el sistema sonoro. El telémetro láser, incorporado en el robot Maggie, proporciona una nube de puntos que

corresponden con la distancia entre los objetos alrededor del robot. Usando esta información, el robot puede seguir la nube de puntos que concuerdan con las piernas del usuario. La distancia y orientación exacta es proporcionada por el sistema de diálogo, en base a las reglas proxémicas vistas en la Fig. 7.7. Observando el comportamiento humano durante una interacción natural por voz, el proceso seguido es muy similar. En primer lugar, se usa el sistema auditivo para aproximadamente localizar la orientación de la fuente sonora (al interlocutor), y el robot se gira hacia dicha orientación. Una vez que el interlocutor queda dentro de nuestro campo de visión, el sistema de visión es usado para determinar si la distancia y la orientación es correcta para mantener la interacción con ese interlocutor concreto. En nuestra opinión, no es necesario tener un complejo y caro sistema de localización basado únicamente en información sonora, debido a que se logra una gran precisión y adaptación integrando el sistema dentro de un sistema de diálogo multimodal.

7.4.3. Integración de la habilidad de localización de usuarios dentro de RDS

Como se ha venido comentado, el sistema de localización de usuarios, se engloba dentro de un completo y complejo sistema de diálogo multimodal que controla el flujo de diálogo, y por lo tanto la interacción entre el humano y el robot. Este diálogo controla un gran número de aspectos a tener en cuenta durante el diálogo, como son el reconocimiento de voz, la síntesis de voz, la generación de gestos que complementan el diálogo (como afirmaciones, negaciones, mirar a los ojos, etc), el reconocimiento de emociones, etc. Uno de los aspectos que también controla y tiene mucho que ver en la consecución de diálogos naturales, es el concerniente a la proxémica entre el propio robot y el usuario.

El sistema de localización sonora descrito en la sección anterior se materializa dentro de la arquitectura de control en el módulo de "localización del usuario". Se puede ver en la figura 7.10 el sistema de diálogo completo y dicho módulo. Este módulo recibe no solo entrada sensorial auditiva, como se ha descrito anteriormente, sino que mediante información visual y de distancia (gracias al telémetro láser infrarrojo), es capaz de realizar fusión multimodal que logra mayor precisión de localización que únicamente mediante el uso de entrada de información sensorial auditiva.

Esta información procesada y fusionada por el módulo de localización sonora, es entregada al dialogo, pasando previamente por otro tipo de fusión multimodal de mayor nivel de abstracción, que organiza toda la información recibida por el resto de módulos en un "macro-paquete" de información procesada, que formalmente se envía al diálogo en un fichero de texto XML, que conceptualmente corresponde con lo que se conoce como actos comunicativos [Falb et al., 2006, Falb et al., 2007, Zaslavsky, 2003].

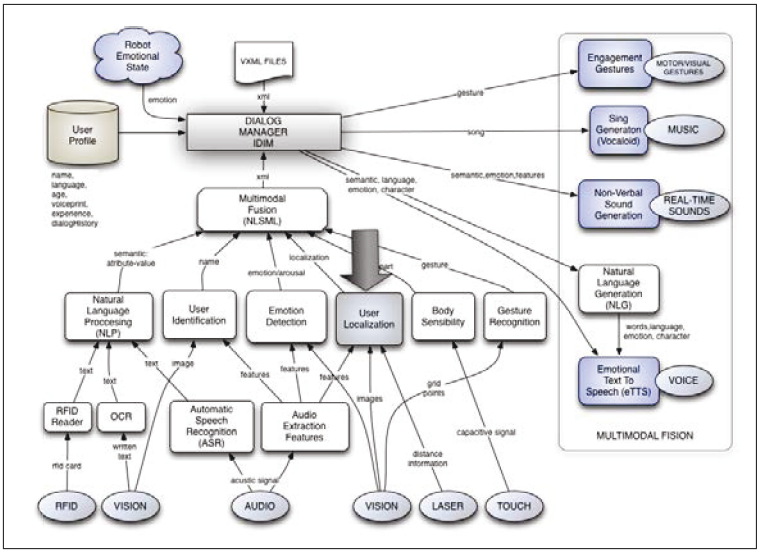


Figura 7.10: Localización de usuarios en el sistema de diálogo multimodal

Es el gestor de diálogo, el que con dicha entrada de información sensorial preprocesada, más el propio perfil del usuario (edad, idioma, experiencia de uso del sistema, nombre e historia del diálogo), puede tomar decisiones inteligentes concernientes a su ubicación espacial respecto al interlocutor. Para tomar esas "decisiones proxémicas inteligentes" es necesario realizar un estudio de como pueden influir dichas entradas de información a la ubicación espacial del robot. Este estudio es realizado en la siguiente sección.

7.5. Experimentos de localización de usuarios

7.5.1. El módulo de localización de fuente sonora

En esta sección, primero se presentan algunos experimentos realizados sólo con el sistema de localización sonora (sólo usando el sistema de audio) para determinar su grado de confiabilidad. Si la precisión del sistema de localización sonora no es suficientemente buena, es muy difícil para el completo sistema de diálogo multimodal determinar la situación espacial adecuada para mantener la interacción.

La habitación donde se han realizado los experimentos tiene unas dimensiones de 11.40 x 6.20 x 3.30 m. con un nivel de reverberación medio, debido a la falta de mucho mobiliario. Para evaluar el sistema de localización sonora, no ha sido necesario un grupo de usuario especial; por lo tanto, los propios miembros del grupo de desarrollo han probado el sistema. En este caso, el usuario se ha situado a diferentes distancias

del robot Maggie, todas ellas comprendidas entre 0.5 y 3m. No se ha encontrado variaciones significativas de precisión en los resultados cuando se mueve entre este rango de distancias. Esto es debido a que el volumen de captura de audio de los micrófonos incorporados en Maggie están ajustados para percibir con suficiente volumen la voz humana a un tono normal hasta los 3m. Los resultados obtenidos son los siguientes:

- Ángulo medio de error en la localización sonora: 23.72º
- Desviación estándar en la localización sonora: 25.82º

Estos valores parecen no suficientemente precisos, pero se debe tener en cuenta que el sistema completo usa fusión multimodal, no sólo el sistema auditivo. La mayor fuente de perdida de precisión en entornos reales es la apariencia de ruidos indeseados, como por ejemplo: el propio sonido por el robot cuando se mueve o habla, o inclusive el sonido de sus propios ventiladores. Para disminuir la incidencia de estos problemas se están desarrollando de algoritmos software (o hardware) de cancelación activa del ruido y cancelación del eco. Otro problema que puede aparecer, es cuando los micrófonos usados no son lo suficientemente "buenos" o las diferencias entre los niveles de captura de audio es demasiado elevado.

7.5.2. El módulo completo de localización de usuarios multimodal

Para probar la utilidad del módulo de localización de usuarios en el sistema de diálogo, se han realizado algunos experimentos, en los que se ha intentado comprobar si el robot se desplaza correctamente acercándose o alejándose del usuario con el que interactúa. Para ello, se ha dividido el espacio alrededor del robot en 4 sectores, (ver Fig. 7.11). En cada una de las zonas se ha situado un usuario diferente a unas distancias de entre 0.5 y 3m. Los usuarios fueron cuatro miembros del departamento ya registrados en su uso con el robot Maggie y por lo tanto con sus perfiles de usuario conocidos por el robot. De edad comprendidas entre 25 y 30 años, tres hombres y una mujer, con valores de experiencia de entre 2 y 150 iteraciones.

El diálogo comienza cuando el usuario saluda al robot. En ese momento el robot detecta aproximadamente la orientación de la que proviene el sonido de la voz humana, se gira el ángulo correcto y se mueve hacia el usuario, manteniendo cierta distancia decidida por el gestor del diálogo. La distancia exacta depende de los estudios proxémicos realizados con el robot Maggie. Durante la interacción humano-robot, el sistema de diálogo comprueba periódicamente la localización del usuario y si la distancia de interacción varia considerablemente (más de 0.5 metros sobre la calculada distancia de interacción ideal), el robot se mueve de nuevo, para colocarse en la posición adecuada. Notar que el usuario puede cambiar su posición durante la

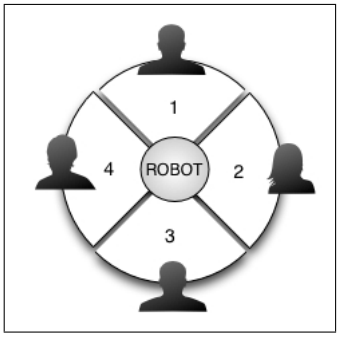


Figura 7.11: Áreas de localización

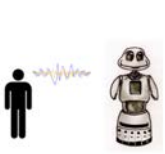
interacción, pero el robot sólo cambiara su posición si la distancia de interacción varia considerablemente. Este proceso es repetido para cada uno de los cuatro usuarios colocados en cada zona. Una descripción intuitiva del proceso puede verse en la Fig. 7.12.

El módulo de localización de usuarios (la localización sonora y la aproximación usando telémetro láser) logra tasas de acierto del 87 %. Eso significa que, si hay un usuario en cada zona, y uno de ellos comienza a hablar, el 87 % de las veces Maggie se gira y se mueve hacia la correcta zona de interacción (con una desviación estándar del 12 %) y mantiene la adecuada distancia de interacción. Estos resultados pueden ser generalizados para otro grupo de usuarios, incluidos niños, de acuerdo a las reglas proxémicas observadas descritas en 7.3.

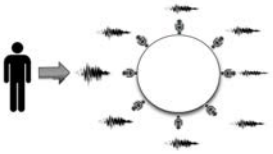
Los errores son debidos principalmente debidos a dos factores: fallos en el módulo de localización de la fuente sonora (sistema auditivo) y/o errores al seguir las piernas del usuario, debido a que algunas veces la nube de puntos que se corresponde con las piernas del usuario es perdida o confundida con otro objeto más cercano de formar similar. Si el error del sistema de localización sonora no es muy alto, este puede ser corregido por el sistema basado en láser, debido a que es capaz de seguir al usuario aunque no esté exactamente colocado centrado enfrente del robot. Por lo tanto, la mayor fuente de errores es confundir objetos cercanos al robot, con las piernas del usuario. Actualmente, se está trabajando en estos problemas, y es de esperar que próximamente el sistema sea complementado con el uso de cámaras esteroscópicas como la Kinect, que pueden facilitar la tarea.

7.6. Resumen

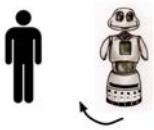
El sistema de localización de usuarios conjuntamente con el estudio proxémico llevado a cabo con usuarios reales interactuando con Maggie, ha proporcionado al sistema de diálogo la facultad de colocar espacialmente al robot a una distancia



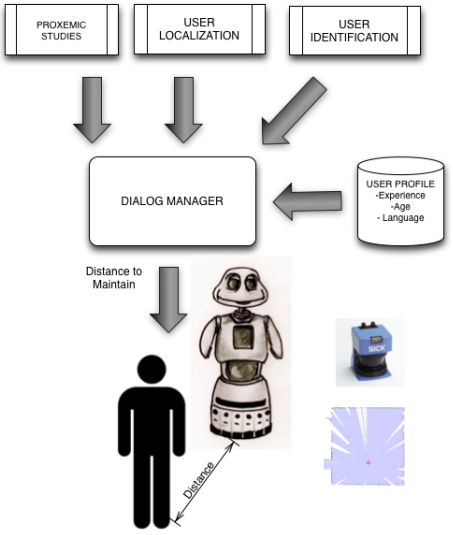
(a) Usuario hablando



(b) Localización sonora



(c) El robot se gira



(d) El robot mantiene la distancia adecuada

Figura 7.12: (**a**) El usuario comienza a hablar al robot Maggie; (**b**) El sistema de localización de usuarios, usando la señal sonora recibida a través de los micrófonos situados en la base del robot, determina el ángulo en que se encuentra el usuario; (**c**) Maggie se gira dicho ángulo; (**d**) El gestor de diálogo basándose en las observaciones realizadas en los estudios proxémicos (ver Fig. 7.7), las medidas obtenidas por el láser, y el perfil del usuario (cargado después de identificar al usuario) da las ordenes a los motores de la base para mantener la distancia adecuada con el usuario.

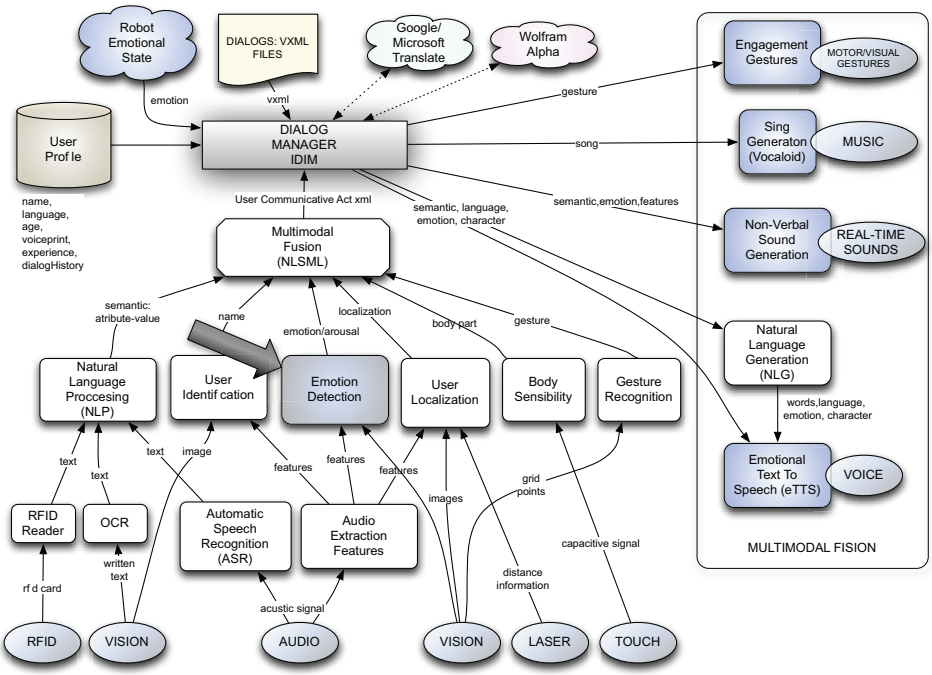
y ángulo respecto al usuario, en el que la interacción se puede realizar de manera natural y satisfactoria. El sistema de diálogo es capaz de adaptar la posición más adecuada para cada situación comunicativa. Ambas tareas descritas en este capítulo son necesarias para lograr ese objetivo: módulo de localización de usuarios y estudio proxémico.

Para la localización del usuario, el robot primero calcula la posición del usuario usando la información acústica que le llega, haciendo uso de 8 micrófonos integrados en la base circular del robot. La distancia de interacción adecuada es determinada de acuerdo a las reglas extraídas en el estudio proxémico, dependiendo del tipo de usuario, su edad, experiencia, etc. Una vez que el robot se ha girado hacía el usuario, el telémetro láser es usado para determinar la distancia al usuario y mantener la distancia correcta.

Recientemente se esta trabajando en una mayor integración sensorial, fusionando información proporcionada por sistemas de visión estéreo, para mayor precisión y robustez del sistema completo. Además, sería interesante incluir mayor personalización de la distancia de interacción (espacio personal). Por ejemplo, si el usuario se siente disconforme con la distancia de interacción calculada por el sistema de diálogo, esta podría ser variada por el propio usuario mediante interacción natural con diálogos.

CAPÍTULO 8

Sistema de detección y gestión de emociones



"Preocúpese hasta de los pequeños detalles."— Steve Jobs

8.1. Introducción

El principal objetivo de los sistemas de diálogo automáticos es conseguir una interacción natural entre el robot y el humano, similar a la que se produce entre las personas. Esto eliminaría la necesidad de artefactos como teclado y ratón en favor de formas de interacción mas intuitivas accesibles por usuarios no expertos y/o con discapacidades. Sin embargo, en los sistemas de hoy en día, la interacción hombre-máquina, aún siendo multimodal, todavía no es comparable a la interacción o diálogo humano. Esto se debe, entre otras cosas, a que la interacción entre personas involucra no sólo el intercambio de contenido explícito, como es la frase literal transmitida mediante voz, sino también de información implícita, como por ejemplo la información acerca del estado emocional del interlocutor. Por lo tanto, una forma de mejorar los sistemas de diálogos sería incluir este tipo de información.

A los sistemas que hacen uso de información emocional se les engloba dentro del campo conocido como "computación afectiva" [Picard, 2000]. Abarca el reconocimiento, gestión, y generación de emociones. En este capítulo se presta atención al proceso de detección de las emociones que expresa el usuario. Su gestión es llevada a cabo por el propio diálogo y la generación es tenida en cuenta por el sistema de síntesis de voz. Recordar, que en cada turno del diálogo se intercambia información, que se agrupa en actos comunicativos, es en cada uno de estos turnos donde se debe añadir, siempre y cuando sea posible, la información relativa a la emoción con la que se expresa el usuario.

Con la inclusión de un módulo específico de detección de emociones, integrado en el sistema de interacción propuesto, se pretende mejorar la interacción, dotándola de mayor naturalidad y "enganche"[1]. Con la detección de la emoción del usuario, el diálogo puede mejorar la adaptación al usuario y a la situación comunicativa concreta. Con ello, se pretende prevenir estados de falta de entendimiento entre ambas partes, que pueden derivar en aburrimiento, desidia, y en último lugar al fracaso del diálogo. De esta forma, el propio sistema de diálogo puede tomar la iniciativa para intentar variar el estado del ánimo del usuario.

En este capítulo se describe e implementa un sistema de detección de emociones multimodal, aplicado a la interacción humano-robot, dentro del sistema de interacción propuesto. Otros trabajos han presentado sistemas de detección de emociones, pero ninguno de ellos integrado dentro de un sistema que gestione la interacción entre robots sociales y humanos. Además, se aporta a la comunidad científica la herramienta GEVA (Gender and Emotion Voice Analysis), desarrollada expresamente en este trabajo, para la detección de la actividad de voz, emociones, y género.

El capítulo se estructura como sigue, en la sección 8.2 se enumeran una serie de

[1]Para entender el concepto de enganche en interacción humano-robot se recomienda leer [Rich & Ponsler, 2010]

cuestiones que tienen que ver con la detección de emociones de manera automática. En la siguiente sección 8.3, se describe cómo el sistema propuesto tiene en cuenta las cuestiones que se han presentado en la sección anterior. Posteriormente, en la sección 8.4 se presenta la herramienta GEVA para la detección de emociones mediante análisis de la voz. En la siguiente sección 8.5, se muestra el sistema de detección de emociones mediante análisis del rostro, este componente se le ha denominado GEFA. En la sección 8.6 se detalla como se fusiona la información multimodal de GEVA y GEFA para tomar una decisión sobre la emoción predominante y como se integra esta con el sistema de interacción RDS. En la sección 8.7 se presentan los experimentos realizados y se analizan. Se finaliza, en la sección 8.8, con las conclusiones obtenidas.

8.2. Trabajo relacionado

En el campo de investigación de detección de emociones aplicado a la interacción humano-robot, se deben tener en cuenta varias cuestiones que se enumeran y describen a continuación:

8.2.1. Las emociones a detectar

De acuerdo con las referencias consultadas, actualmente, para clasificar o etiquetar una emoción existen principalmente dos aproximaciones. La primera consiste en un conjunto discreto de emociones básicas, de las cuales se pueden generar otras secundarias [Arnold, 1960]. Siguiendo esta aproximación, Ekman establece un conjunto de 6 emociones básicas (alegría, repugnancia/asco, ira/enfado,miedo, sorpresa, y tristeza) [Ekman & R.J, 1994, Ekman et al., 1972].

La segunda aproximación consiste en medir y contextualizar las emociones de acuerdo a diferentes dimensiones. De esta manera, una emoción puede ser entendida como puntos o áreas en el espacio definidos por esas dimensiones. En los trabajos de Plutchik y Cowie [Plutchik, 1980][Cowie & Douglas-Cowie, 2000] se establecen las dimensiones de *activación*, entendida como la predisposición de la persona para tomar alguna acción de acuerdo a su estado emocional, y *evaluación*, que refleja una valoración global del sentimiento positivo o negativo asociado al estado emocional. En los trabajos de Bradley [Bradley & Lang, 2000, Bradley & Lang, 1994] se usa un modelo excitación-placer (*arousal-pleasure*), la primera dimensión es la excitación (valencia), que explica el deseo, y la segunda explica la actividad fisiológica relacionada con el estado afectivo .

El uso de un espacio continuo para clasificar un estado emocional puede verse como un paso intermedio antes de determinar el estado emocional al que corresponde ese punto, pero esto no siempre tiene por que ser así. En nuestro caso, el sistema de detección de emociones podría trabajar única y exclusivamente con estos valores

continuos y entregárselos al gestor del diálogo para que los maneje a su antojo, sin discretizar [2] la salida en una emoción concreta.

8.2.2. Canales utilizados para la detección emocional

En la literatura se pueden encontrar varios trabajos relativos a la detección de emociones automática, siendo los más comunes aquellos relacionados con el análisis mediante visión artificial del rostro humano. Otros trabajos, analizan algo menos común, que tiene que ver con el análisis de la voz del usuario. Finalmente, unos pocos describen sistema multimodales para la detección de emociones. Sin embargo, hasta ahora no se ha encontrado ningún trabajo que describa e implemente un sistema de detección de emociones aplicado a un sistema de diálogo-interacción humano-robot.

De Silva [De Silva et al., 2007] concluyó que algunas emociones son mejor identificadas por voz, como tristeza y miedo, mientras que otras eran mejor identificadas mediante vídeo, como el enfado y la felicidad. Además, Chen [Chen et al., 1998] mostró que esas dos modalidades dan información complementaria, con el argumento de que el rendimiento del sistema aumenta cuando ambas modalidades son consideradas en conjunto. Esta idea parece ser corroborada en el trabajo de Tu [Tu & Yu, 2012], que proclama que la tasa de detección de emociones correctamente por cada canal sensitivo alcanza valores del 60 % (voz) y el 57 % (imagen), mientras que un único clasificador multimodal logra tasas de acierto del 72 %. Otros trabajos, únicamente basados en el análisis de la voz, proclaman tasas de acierto de entre el 65 % y el 88 % de acierto [Cowie et al., 2001, Pantic et al., 2005, Roy & Pentland, 1996].

Por otro lado, Yoshitomi [Yoshitomi et al., 2000] complementa la extracción de características del rostro con el añadido de un nuevo canal de entrada de información no intrusivo, como es el de la imagen de temperatura obtenida por una cámara infrarroja. En ese trabajo los tres canales son fusionados a nivel de toma de decisión, por lo que existe un clasificador para cada canal. Finalmente, se afirma mayor porcentaje de reconocimiento con el sistema general en comparación con cada uno de los subsistemas de detección por separado. La tasa de acierto es de un 60 % en el canal sonoro, 56 % en el visual, 48 % con IR, y de un 85 % en el sistema global combinando los tres canales anteriores.

En las dos últimas décadas, se han realizado estudios usando interfaces cerebro-máquina para la detección de emociones. Se basan en el estudio de la actividad neuronal de los usuarios [Petrantonakis, Panagiotis C. & Hadjileontiadis., 2010, Khalili & Moradi, 2009, Murugappan & Rizon, 2008, Valderrama Cuadros, C. E., Ulloa Villegas, 2012]. La técnica más usada suele ser la del análisis electroencefalográfico (EEG). En estos

[2] Convertir algo continuo en un conjunto de valores.

trabajos se explica cómo trabajar con las señales neuronales y cómo extraer información de ellas utilizando la *Transformada Discreta Wavelet* (DWT)[3]. También se mencionan los métodos que se usan para la clasificación y caracterización, que son los mismos que los usados en otros canales sensitivos. En este caso, la tasa de acierto se sitúa entre el 55 % y el 76 %.

Todos estos trabajos muestran sistemas de detección de emociones que son presentados sin tener en cuenta su uso en una aplicación práctica real. Por lo tanto, dichos sistemas no están acoplados e integrados en un sistema general que sea susceptible de mejorar sus prestaciones gracias a la inclusión de esta información emocional.

8.2.3. Nivel al que se fusiona la información de cada canal

Todo sistema de detección automático de emociones presenta dos fases fundamentales: la fase de extracción de características y la fase de clasificación de las mismas en emociones. Dado que se está hablando de un sistema multimodal para la detección de emociones, es necesario determinar en que momento hacer la fusión de la información extraída por los canales involucrados.

Por un lado cabe la posibilidad de tener un único clasificador que reciba como entrada las características extraídas por cada canal y tome la decisión de la emoción expresada[4] mediante algún tipo de algoritmo. En este caso, se dice que la fusión se hace a "nivel de extracción de características".

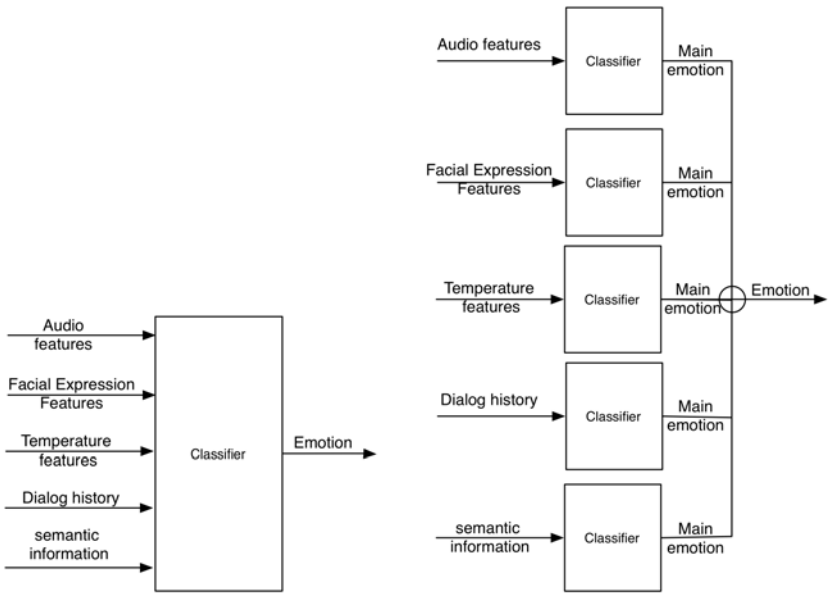
Por otro lado, existe la posibilidad de que cada canal determine de manera aislada la emoción que se percibe exclusivamente usando la información percibida por ese canal. En este último caso, es necesario una regla de decisión que tenga en cuenta la emoción percibida por cada canal, y la confianza que se tiene en dicha detección, para determinar la emoción final predominante. En este último caso la fusión se hace a "nivel de decisión". En este sentido, el psicólogo, Mehrabian dio la siguiente fórmula para establecer "pesos" en la detección de emociones: significado oración (semántica) 7 %, entonación de la oración 38 %, rostro 55 %.

En la Fig. 8.1 se ilustra, a modo de ejemplo, estas dos aproximaciones usando varios canales sensitivos.

Truong [Truong et al., 2007] realiza un resumen comparativo de varios de los experimentos realizados por otros autores, siendo los principales: Chen, Sebe, y De Silva. En el experimento de Chen [Busso et al., 2004], se muestra que la tasa de acierto del canal auditivo es del 75 %, la visual del 69 %, y la combinación de ambas a nivel de extracción de características (un único clasificador para ambos canales) llega al 97 %. En el experimento de Sebe [Sebe et al., 2006] la tasa de acierto para el canal

[3]Es la transformada que permite procesar señales no estacionarias obtenidas en el EEG

[4]de una manera similar a como en el cerebro humano se conectan la información que se recibe tanto por la vista, el olfato, el oído, etc.



(a) Un único clasificador: fusión a nivel de extracción de características

(b) Un clasificador para cada canal: fusión a nivel de toma de decisión

Figura 8.1: Nivel al que se realiza la fusión: a nivel de decisión o a nivel de extracción de las características

de audio es del 45 %, 56 % para el visual, y del 89 % de la combinación de ambos a nivel de extracción de características. En cambio, en el experimento de De Silva [De Silva et al., 2007], en el que la fusión se realiza a nivel de decisión, se afirma que obtiene una tasa de acierto del 72 %. Estos trabajos parecen demostrar dos cosas: que la fusión de información de varios canales mejora notablemente la tasa de acierto del reconocimiento de emociones, y que la fusión a nivel de extracción de características logra tasas de acierto mayores que a nivel de decisión[5].

8.2.4. Tipos de entrenamiento del sistema

Una vez que se decide el uso de un clasificador para cada canal, o de un único clasificador global para todos los canales, es necesario determinar si se construyen/entrenan para cada usuario, para cada grupo de usuario (por ejemplo que comparten unos mismos rasgos culturales como el idioma), o en general, para cualquier tipo de usuario. Además, hay que determinar si se entrena durante la propia interacción natural con el sistema (*online*) o se realiza previamente a su uso (*offline*). Por otro lado, también es necesario tener en cuenta si el entrenamiento se realiza forzando la expresión de emociones de manera no natural, o se realiza de manera espontánea.

8.3. Características del sistema multimodal de detección de emociones propuesto

En esta sección se describe el sistema que se ha desarrollado e integrado dentro del sistema de interacción general (se puede ver un esquema general en la Fig. 8.2). Por ello, se va a ir analizando punto por punto como se han resuelto los desafíos planteados en la sección anterior.

8.3.1. Representación de las emociones a detectar

La primera decisión a adoptar es cómo clasificar las emociones percibidas. Cabe la posibilidad de intentar clasificar directamente la emoción siguiendo la aproximación de Ekman [Ekman et al., 1972], en un conjunto reducido de posibles emociones. La otra opción, consiste en intentar percibir el nivel de activación y atracción para, si así se quiere, determinar el área del espacio afectivo correspondiente y la emoción relativa a esa coordenada.

[5]Truong comenta que las comparaciones no son del todo rigurosas puesto que el conjunto de emociones es diferente, entre otras cosas.

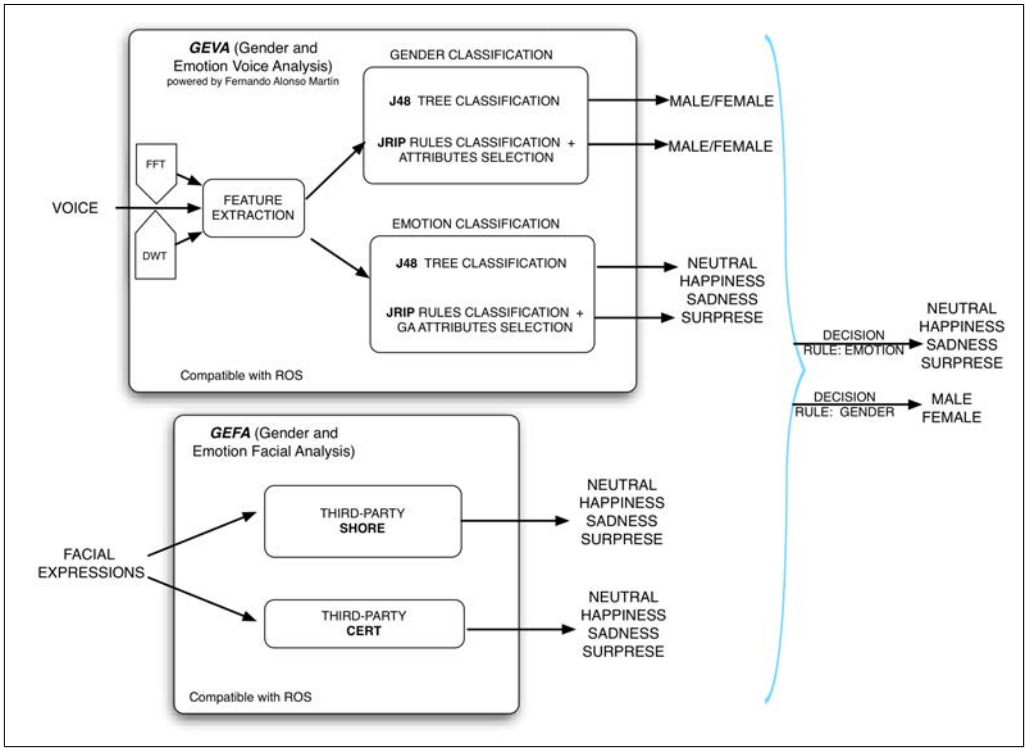


Figura 8.2: Sistema de detección de emociones multimodal implementado. Esta formado por dos componentes, uno basado en el análisis de la voz del usuario (GEVA) y otro basado en el análisis del rostro del usuario (GEFA). Ambas salidas son fusionadas mediante una regla que decide cual es la emoción predominante

Se ha optado por el primer modelo, que consiste en un conjunto de emociones finitas y un valor de grado de confianza para cada una de esas emociones. Está aproximación se elige por varias razones:

- Facilidad para adaptar la implementación de nuevos diálogos de interacción. Dado que el sistema de diálogo/interacción debe usar la emoción percibida del usuario, y puesto que esto se hace mediante la especificación del diálogo en forma de fichero XML (que sigue el estándar VoiceXML [6] y que interpreta el gestor de diálogo[7]). Se hace más sencillo e intuitivo escribir sentencias como: *"si el usuario está feliz le propongo jugar a un juego"*, en vez de: *"si el nivel de activación > 50 y el nivel de evaluación es > 90 entonces le propongo jugar a un juego"*.

- Facilidad para unificar la información proveniente de los distintos canales/modos usados para detectar la emoción. Al discretizar la emoción en un conjunto finito de estados emocionales se hace más sencilla la integración de los distintos modos.

El conjunto de emociones que finalmente se manejan en nuestro sistema completo es de cuatro: *neutro, felicidad, tristeza y sorpresa*. La elección de este conjunto concreto de emociones obedece a las siguientes razones:

- No se quiere un conjunto extenso de emociones, ya que a mayor número posible de grupos a clasificar, el nivel de acierto disminuye.

- Para la interacción humano-robot están son las que se han considerado más relevantes a la hora de personalizar la interacción.

- Son emociones relativamente fáciles de discernir y que aparecen en los distintos modos seleccionados para detectar la emoción. Algunas emociones definidas por Ekman se han descartado: miedo, asco e ira. Este descarte se ha se ha debido principalmente a que las herramientas involucradas, que usan información audiovisual, presentan importantes dificultades para su detección.

Por otro lado, como algunos autores comentan [Litman & Forbes-Riley, 2004], frecuentemente el porcentaje de discurso neutro frente al emotivo está muy desequilibrado, siendo mucho mayor el neutro que el emotivo. Por ello, en el ámbito de la detección de emociones automáticas, el estado de neutralidad o tranquilidad es considerado como una emoción más a detectar por el sistema automático.

[6] `http://en.wikipedia.org/wiki/VoiceXML`

[7] Basado en la herramienta de relleno de huecos de información Bladeware: `http://sourceforge.net/projects/bladeware-vxml/`

8.3.2. Canales utilizados para la detección de emociones

Al igual que ocurre en la interacción entre humanos, existen diversos canales a tener en cuenta a la hora de percibir con precisión la emoción expresada por el usuario durante la conversación. En este trabajo, los canales usados para la detección de emociones son el sonoro (análisis de la voz del usuario) y el visual (análisis de las expresiones faciales del rostro). Como se verá se han integrado las librerías (SHORE y CERT) que analizan mediante visión el rostro humano dentro de un módulo denominado GEFA (*Gender and Emotion Facial Analysis*). Por otro lado se ha desarrollado el módulo de análisis de la voz del usuario, al que se ha denominado **GEVA** (Gender[8] and Emotion Voice Analysis).

Por el momento, el uso de técnicas invasivas ha sido descartado para este trabajo dado que la naturalidad y la satisfacción en la interacción podría verse mermada. El uso de análisis de imágenes de temperatura corporal no se ha tenido en cuenta al no disponer del sensor necesario (por tamaño y por precio).

8.3.3. Nivel al que se fusiona la información de cada canal

Tal y como se describe en la sección 8.2.3, existen diversas metodologías para realizar la integración multimodal, principalmente existen dos alternativas: un clasificador para cada modo y posteriormente una regla de decisión que fusione las salidas de cada modo (fusión a nivel de decisión), o un único clasificador que reciba las características extraídas por cada modo (fusión a nivel de extracción de características).

Parece muy difícil hacer comparaciones entre estas dos metodologías, ya que el conjunto de estados, usuarios, calidad del sonido e imagen es muy diferente en cada caso. Para este trabajo, *se ha adoptado la primera aproximación que consiste en tener un clasificador para cada canal, y finalmente una regla de decisión, que determina qué emoción prevalece en ambos clasificadores*. La razones fundamentales para seguir esta metodología han sido:

1. Histórica: inicialmente se desarrollo un clasificador usando sólo voz, dado que el campo del sonido nos es más familiar, que posteriormente se ha complementado con el visual.

2. Facilidad de desarrollo y evaluación: se hace más sencillo "depurar" cada sistema por separado que un único sistema combinado, dado que hay menos factores involucrados. Si se obtiene una alta tasa de acierto por cada canal, un sistema combinado debería obtener también una alta tasa de acierto.

[8] Aunque esta fuera del ámbito del trabajo expuesto, el software desarrollado es capaz de detectar, además de la emoción, el género del hablante.

3. Facilidad de integración: dado que se hacen uso de alguna herramienta comercial, que incluye extracción de características y clasificación (de manera interna), no es posible leer la extracción de características.

8.3.4. Tipos de entrenamiento del sistema

En este trabajo el entrenamiento de los clasificadores usados para cada canal se ha realizado *offline*, en base a ejemplos previamente obtenidos y etiquetados. Su funcionamiento es universal, es decir, sin necesidad de entrenamiento para cada usuario o cultura (idioma o dialecto). Además, para el entrenamiento y para la validación (experimentos) se incluyeron (en el caso del canal de voz) expresiones de emociones espontáneas y forzadas.

Las expresiones espontáneas durante el entrenamiento fueron recogidas mediante el análisis de vídeos de conversaciones naturales en internet. Por otro lado, las forzadas se obtuvieron de entrevistas con usuarios a los que se les pidió que actuaran simulando una determinada emoción, así como mediante vídeos de series y películas.

Para la fase de validación (experimentación) se probó con interacciones naturales, en las que los usuarios se dirigían al robot sin la presencia de ningún supervisor. También se realizaron experimentos supervisados en los que se pedía al usuario que simulasen diversas emociones.

8.4. El sistema de detección por voz propuesto: GEVA

En la literatura se pueden encontrar evidencias sobre cómo el cambio de emoción en el interlocutor repercute en su tono de voz [Cowie, R., Douglas-Cowie, E., & Romano et al., 1999]. Estos cambios se reflejan en variaciones prosódicas en la voz del interlocutor. Sin embargo, los usuarios raramente alcanzan una "emoción pura", entendiendo esto como que casi nunca se expresa una emoción de completa felicidad, o completa tristeza sin ser mezclada con otras emociones. En ese mismo trabajo, se muestra como cambios en la "evaluación" y "activación" de una emoción se ven directamente asociados a cambios prosódicos en el discurso verbal.

Para la detección de emociones mediante el canal sonoro, son fundamentales dos pasos: el primero es el de la extracción de características sonoras de la locución, y el segundo es el de la construcción de un clasificador que produzca como salida la emoción percibida por ese canal. Para la construcción de dicho clasificador es necesario saber que características se van a tener en cuenta, el conjunto posible de emociones a

clasificar, y un conjunto de locuciones de entrenamiento (que se denominará corpus)[9].

8.4.1. La extracción de características sonoras

Existen varios sistemas abiertos de extracción de características sonoras. En este trabajo se ha experimentado con "OpenSMILE" [Eyben et al., 2010] y su versión con clasificador de emociones: "OpenEAR" [Eyben et al., 2009]. Otro entorno de extracción de características sonoras es "PRAAT" [Boersma, 2002], pero se ha entendido que no es tan adecuado para extraer características del audio en tiempo real. La empresa española Biloop [Ramírez, 2009] ha desarrollado su propio software/hardware para la detección y clasificación de emociones, con especial interés y aplicación en los bebés. Verbio Speech Analytics[10] proporciona los mecanismo para el análisis del audio y clasificación en emociones (enfado, contento y neutral).

Para la extracción de características sonoras se ha usado un lenguaje de análisis y de generación sonora llamado Chuck[11]. La elección de Chuck para esta tarea se ha basado en su capacidad para trabajar en tiempo real con la entrada de audio, su simplicidad y elegancia de programación utilizando los analizadores de audio incorporados. Por lo tanto, Chuck permite de manera fácil, rápida, y elegante conectar la extracción de características de la entrada de sonido con la generación, también en tiempo real, de sonidos no verbales[12]. Se puede comprobar la versatilidad y potencia de Chuck para análisis y síntesis de Chuck mediante la herramienta Wekinator [Fiebrink & Cook, 2010].

En este trabajo se ha desarrollado un módulo completo e independiente del resto para realizar esta tarea. Este módulo es de suma importancia puesto que existen otros del sistema RDS que consultan la salida ofrecida por este módulo para realizar otras tareas: localización e identificación de usuarios, detección de actividad de voz, detección del nivel de arousal, etc.

Este módulo de extracción de características sonoras está en continua ejecución y analiza continuamente el sonido, mientras se perciba voz del usuario, extrayendo ciertos valores instantáneos a los que se denominaran, a partir de ahora, estadísticos. Una vez que el usuario ha dejado de hablar, y por lo tanto se considera que su turno ha concluido (acto comunicativo), se calculan los valores máximos, mínimos y promedios de los estadísticos anteriores. Son los siguientes:

[9]La construcción del clasificador necesita de un proceso de aprendizaje o entrenamiento, que se hace mediante muestras de voz etiquetadas en una emoción concreta

[10]`http://www.verbio.com/webverbio3/es/tecnologia/verbio-tts/52-verbio-speech-analytics.html`

[11]http://chuck.cs.princeton.edu/

[12]Actualmente Chuck está siendo usado por millones de personas mediante aplicaciones para Android y iOS, bajo la empresa de su creador: SMULE.

1. El **pitch** se refiere a la altura de los sonidos que percibe el oído humano[13], depende del número de vibraciones por segundo producidas por las cuerdas vocales. El cálculo del pitch, también conocido en algunos casos como frecuencia fundamental [14] (aunque no es exactamente lo mismo), no está incluido en los componentes de análisis de Chuck, por lo que se ha tenido que implementar.

   En la literatura se han propuesto numerosos algoritmos para el cálculo del pitch, y hoy en día, continua siendo un campo de investigación vigente. Nuestro algoritmo trabaja en tres dominios: el del tiempo (usando la autocorrelación), el de la frecuencia (usando la transformada rápida de Fourier), y el del tiempo-frecuencia (usando la transformada de Wavelet).

2. El **flux** o la variación temporal del espectro, indica si hay grandes variaciones de amplitud en el espectro (dominio de la frecuencia). Valores cercanos a 0 indican que los valores obtenidos de amplitud, en todos los rangos de frecuencia son muy similares. Valores cercanos a 1 indican que hay profundas variaciones en las amplitudes del espectro, mostrando un espectro muy dentado. El estadístico flux nos habla de si la señal se concentra entorno a una amplitud promedio, o hay muchas variaciones en el volumen de dicha señal.

3. El estadístico **rolloff-95** corresponde al valor de frecuencia para la que el 95 % de la energía de la señal está contenida en esa frecuencia.

4. El estadístico **centroide** representa la mediana del espectro de una señal. Es decir hacia que frecuencia se aproxima más la señal analizada. Es usado a menudo para calcular el tono o la brillantez de un sonido o voz. Nuestro aparato fonador produce valores muy diferentes a los que produce el sonido generado por un violín o una flauta.

5. El estadístico **zero crossing rate** indica la cantidad de veces que la señal atraviesa el zero (el eje x), típicamente el ruido de fondo cruza muchas más veces el eje x que un sonido verbal.

6. El cálculo del estadístico **SNR** permite relacionar el volumen (RMS) de la señal de voz con el volumen de la señal de ruido. Para ello, se calcula el volumen medio del ruido cuando no se detecta actividad de voz. También se calcula el volumen medio de la señal recibida cuando se detecta actividad de voz. Finalmente, SNR divide el volumen medio de la señal recibida cuando se detecta actividad de voz

[13]Entendiendo altura como frecuencia, no volumen

[14]La frecuencia fundamental para una señal armónica (varios sonidos oscilando al mismo tiempo), es la frecuencia mínimo común múltiplo de todas ellas. Es decir, cualquiera de los armónicos que forman la señal se pueden conseguir por múltiplos de esa frecuencia base.

entre el volumen medio de la señal recibida cuando no se detecta actividad de voz. SNR permite reflejar con mayor fidelidad el volumen de la voz del usuario, ya que se hace respecto al ruido ambiente.

7. Mediante el estadístico **ritmo comunicativo** se trata de obtener información sobre el mayor o menor ritmo comunicativo del usuario. Para ello, se obtiene el número de palabras pronunciadas por minuto. Cada palabra se encuentra separada de la anterior por un breve instante de silencio que es identificado. Se multiplica el número de silencios detectados por 60 y se divide entre el número de segundos que ha durado ese turno de voz.

8.4.2. La clasificación por voz

Una vez que se han extraído las características que permiten caracterizar la locución de voz del usuario, es necesario clasificar a qué emoción corresponden. A continuación se ven las opciones posibles para entrenar/construir el clasificador, en base a: si se hace en interacciones reales (online), o en interacciones previas (offline); si se simula o no la emoción; además de determinar quién es el encargado de etiquetar la emoción expresada:

- **Offline**: esta aproximación consiste en entrenar el sistema de detección de emociones automático previamente a su uso final dentro del sistema de interacción general. Existen, al menos, dos posibilidades:

  1. Un único clasificador universal (para todo tipo de usuario) entrenado previamente, utilizando ejemplos con locuciones de todo tipo de usuarios (edades, idiomas, culturas, y géneros). Sin embargo, en la literatura se encuentran algunos trabajos que afirman que es más preciso un clasificador para cada locutor que un único clasificador universal [Callejas & Lopezcozar, 2008, Eyben et al., 2010].
  2. Un clasificador para cada idioma posible. Con esto se supone que se conseguiría una mejora importante en la tasa de acierto. En este caso, es el propio diálogo el que informa al módulo de detección de emociones del idioma en el que se está dialogando. Otra pequeña variación consiste en el uso de dos clasificadores por idioma, uno para masculino y otro para femenino.

- **Online**: construir el clasificador en interacciones reales con el usuario usando el sistema de interacción general (en nuestro caso es el sistema RDS), siendo durante la fase de registro con el sistema el momento oportuno. La fase de registro se hace una única vez por usuario en el sistema. Durante esta fase

el sistema aprende características del usuario como su: nombre, idioma, tono de voz, forma de expresar sus emociones, etc. Hay varias posibilidades para la construcción del clasificador durante el registro del usuario con el sistema:

1. El usuario se registra en el sistema y un supervisor externo etiqueta cada locución con la emoción con la que cree que se ha expresado el usuario. Esta aproximación presenta varios problemas: primero, resulta difícil que únicamente durante la fase de registro el usuario se exprese usando todas las emociones posibles (las cuatro establecidas); segundo, probablemente no se obtengan una cantidad suficiente de ejemplos por cada emoción; y tercero, la necesidad de un supervisor que etiquete dichas emociones.
2. Durante la fase de registro, el sistema le pide al usuario que se exprese con cada una de las emociones que el sistema quiere aprender. Se necesita que el usuario simule dichas emociones, tantas veces como sean necesarias, para que el sistema aprenda sus características y construya el clasificador. En este caso, el diálogo actúa como supervisor, sin necesidad de un humano que etiquete las locuciones. Los inconvenientes asociados son que las emociones simuladas no suelen ser tan precisas como las obtenidas en interacciones reales; además, no resulta la manera más natural de interactuar con el sistema puesto que para considerarse, esa fase de registro, como natural debe tratar de imitar la manera que tienen los humanos de conocerse.

- **Mixta**: en esta aproximación, es necesario de un clasificador construido en base a las relaciones relativas entre los estadísticos descritos y no en base a los valores absolutos que pueden tomar (por ej. la felicidad se refleja en un tono de voz con un pitch un 20 % más elevado que la emoción neutra). La construcción de este clasificador se realizaría *offline*. Posteriormente, en interacciones reales (*online*) durante la fase de registro del usuario con el sistema, se obtienen los valores de los estadísticos asociados al tono de voz neutro del usuario. Con estos valores, asociados al tono neutro, se escala el clasificador offline, de tal manera que esas relaciones relativas entre los estadísticos se convierten en reglas con datos numéricos absolutos. Esta aproximación requiere la licencia de suponer que la emoción expresada por el usuario durante la fase de registro es la neutra (esta aproximación es propuesta por [Callejas & Lopezcozar, 2008], ya que lo que le interesa es estudiar las emociones derivadas de la interacción del usuario con el sistema y no tanto las previas. Si el usuario se registra con el sistema usando un tono de voz diferente del neutro, esta aproximación fallaría.

En este trabajo se ha optado por la creación de un clasificador de emociones universal y entrenado *offline*. Para la toma de esta elección se ha considerado que si

no se obtenían resultados suficientemente satisfactorios se adoptaría la aproximación del uso de un clasificador *offline* para cada lenguaje o dialecto. De acuerdo a la elección tomada ha sido necesario el uso de locuciones etiquetadas en una de las cuatro posibles emociones consideradas (felicidad, tristeza, sorpresa, y neutralidad). Para ello, se han usado las siguientes fuentes que proporcionan muestras de voz relativas a varias emociones en diversos idiomas, edad, y género:

- Ejemplos de voz de los propios desarrolladores simulando emociones.
- Entrevistas a compañeros de trabajo en los que se les pide simulación de emociones.
- En interacciones reales (espontáneas) del robot con compañeros de trabajo.
- Mediante entrevistas obtenidas de internet.
- Mediante series de televisión con actores reales obtenidas de internet.
- Mediante audiolibros subidos en internet.
- Mediante bases de datos con corpus de voz etiquetado:
  - Emotional Prosody Speech Database (LDC): con 15 clases de emociones.
  - Berlin Emotional Speech Database (EmoDB): con 7 clases emociones. Ver [Vlasenko & Schuller, 2007b, Vlasenko & Schuller, 2007a].
  - Danish Emotional Speech Corpus (HUMANAINE): con 5 clases emociones. Ver [Schuller et al., 2004, Schuller & Arsic, 2006].
  - FAU Aibo Emotion Corpus[15]: 8.9 horas de voz espontánea de 51 niños, en 11 clases. Ver [Steidl, 2009].

Con este corpus lingüístico, y sometido a nuestro módulo de extracción de características sonoras, se ha construido un fichero con patrones de entrenamiento para Weka [Holmes et al., 1994] con unas 500 entradas de locuciones etiquetadas en emociones. Las muestras recogidas están en formato *arff* (válido para la herramienta Weka).

El programa de aprendizaje automático Weka mencionado, recoge multitud de técnicas de aprendizaje automático que nos permiten realizar la clasificación partiendo de los patrones de entrenamiento. En esta fase de entrenamiento se han obtenido diferentes valores de precisión (tasa de acierto) en la detección correcta de emociones[16], usando diferentes técnicas:

[15] `http://www5.cs.fau.de/de/mitarbeiter/steidl-stefan/fau-aibo-emotion-corpus/`

[16] Estos resultados se han obtenido mediante la técnica de "validación cruzada" durante la fase de entrenamiento del sistema. En experimentos reales la tasa de acierto en sensiblemente menor, puesto que se valida con nuevas muestras, mientras que "validación cruzada" valida con las mismas muestras con las que se entrena el sistema.

1. **Bayesiano**. Red Bayesiana: 68.95 %, Naive Bayes: 65.52 %,
2. **Lógica borrosa**. IBK: 85.65 %, IB1: 82.22 %, LWL: 62.74 %
3. **Reglas**. JRIP: 81.15 %, ConjuntiveRule: 61.88 %, DecisionTable: 70.02 %, ZeroR: 57.17 %, PART: 77.94 %
4. **Árboles de decisión**. J48: 80.51 %, BFTree: 79.01 %, LADTree: 68.95 %, LMT: 79.44 %,

La elección de los clasificadores que se integran finalmente en GEVA se justifica en base a aquellos algoritmos que mayor porcentaje de acierto consiguen usando el método de "valización cruzada" sobre el conjunto de entrenamiento. También se ha tenido en consideración su facilidad de implementar (o integrar) en Chuck.

GEVA proporciona, en tiempo real, las salidas de emoción y género (este último no tenido en cuenta para este trabajo) usando tres métodos para cada uno de ellos, así como los valores de confianza. Los algoritmos de clasificación usados son:

- Árboles de decisión J48[17] usando todos los estadísticos de la voz descritos previamente.
- Reglas de decisión JRIP[18] usando todos los estadísticos de la voz descritos previamente.
- Reglas de decisión JRIP + Algoritmos Genéticos: usando los estadísticos seleccionados mediante algoritmos genéticos de entre el conjunto de los descritos previamente[19].

8.5. El sistema de detección visión del rostro propuesto: GEFA

En numerosos trabajos se hace uso de la información visual, en especial la relativa a la cara del usuario, de manera individual [Bartlett et al., 2003] o complementaria a la voz [Tu & Yu, 2012, Busso et al., 2004]. En el trabajo [Pantie & Rothkrantz, 2000], se presenta un extenso resumen sobre técnicas y trabajos realizados para dicha detección de emociones mediante visión. En ese trabajo se siguen unos pasos muy similares a los realizados en la detección de emociones mediante voz:

[17]Es una implementación que hace Weka del árbol de clasificación C4.5, se usa en minería de datos o aprendizaje automático

[18]Routing Information Protocol, es un algoritmo del tipo vector-distancia usado en técnicas de minería de datos o aprendizaje automático

[19]En lugar de usar todos los estadísticos de la voz, sólo se usan los que aportan significado trascendente, el resto se consideran ruido, ya que no aportan valor a la clasificación

- Detección de la cara: detectar la cara del usuario en el flujo de imagenes.
- Detección de las características de la cara: características como distancia entre ojos, forma de la boca y párpados, etc.
- Clasificación de la emoción: partiendo de las características extraídas se construye un clasificador.

8.5.1. Detección de la cara

La detección de la cara del usuario o su seguimiento en una escena de vídeo continua (*tracking*), ha sido ampliamente estudiado. Existen trabajos, como el de Viola en 2004 ([Viola & Jones, 2004]), que presentan la detección de caras con sistemas automáticos robustos. Además, la conocida librería de visión artificial OpenCV trae incorporada las funciones necesarias para realizar dicha tarea[20]. En recientes trabajos en la librería OpenCV también se han añadido las funciones necesarias para realizar el reconocimiento del género (masculino o femenino) [21].

8.5.2. Extracción de las características del rostro

Una vez que la presencia de la cara ha sido detectada en la escena observada, el siguiente paso es extraer la información sobre la expresión de la cara encontrada de una manera automática. Sobre la manera de representar y extraer la información asociada a dicha expresión en el rostro hay varias aproximaciones computacionales.

Una de las aproximaciones es la realizada en función de puntos de interés, también conocida como aproximación local, y su relación geométrica entre ellos [Kobayashi & Hara, 1997, Padgett & Cottrell, 1997]. En esta aproximación es necesario detectar ciertos puntos asociados al rostro. Sobre estos puntos y la relación entre ellos se extraen las características geométricas oportunas.

En otra aproximación, la cara es representada como una unidad completa (aproximación holística). Una malla 3D con una textura se sitúa sobre la cara reconocida [Terzopoulos & Waters, 1993, Lucey et al., 2006]. Esta malla tridimensional o también conocido como "modelo de aspecto activo", presenta mayor robustez frente al movimiento de la cabeza en tiempo real, así como oclusiones parciales.

Por último, la cara puede ser modelada mediante una aproximación híbrida de los dos métodos anteriores, que tipifica una combinación de las aproximaciones analíticas y holísticas para la representación de la cara. En esta aproximación, un conjunto de

[20] `http://docs.opencv.org/trunk/modules/contrib/doc/facerec/`

[21] `http://docs.opencv.org/trunk/modules/contrib/doc/facerec/tutorial/facerec_gender_classification.html`

puntos de interés de la cara se usan para determinar la posición inicial de la malla que modela la cara [Kearney & McKenzie, 1993].

Independientemente del tipo de modelo de la cara aplicado, la extracción de las características debe hacerse sin perdida de información, o con la mínima perdida posible. Algunos factores hacen esto una tarea compleja, la primera es la presencia de pelo, gafas, etc, que ocluyen determinadas facciones del rostro. Otro problema es la variación en el tamaño y orientación de la cara en las imagenes de entrada. Esto imposibilita una búsqueda de patrones fijos en las imágenes. Finalmente, el ruido en la imagen constituye otro inconveniente.

8.5.3. Clasificación de la expresión del rostro

Después de que la cara y su apariencia hayan sido percibidos, el siguiente paso consiste en analizar la expresión automáticamente. Mientras que el mecanismo humano para la detección de caras es bastante robusto, no sucede lo mismo en la detección de expresiones faciales. Un observador entrenado, puede correctamente identificar expresiones en un 87 % de los casos, valor que varía dependiendo en gran medida de la familiaridad que tenga con la persona a observar [Bassili, 1978].

El Sistema de Codificación de Acciones Faciales (Facial Action Coding System o FACS [Ekman et al., 1978]) es probablemente el estudio mas conocido sobre actividad facial y es ampliamente usado. Constituye una representación desarrollada para permitir a los psicólogos codificar expresiones de la cara partiendo de imágenes estáticas. Está basado en la enumeración de todas las "unidades de acción"(AU) de la cara, concretamente 46, que pueden causar movimientos faciales. La combinación de esas AUs da como resultado un amplio conjunto de expresiones faciales posibles. Por ejemplo, la expresión de sonrisa puede ser considerada como la combinación de "tirar de la comisura de los labios"(AU12 + 13) y/o "apertura de la boca" (AU25 + 27) con "elevar el labio superior"(AU10) y un poco de "profundidad en el surco lateral" (AU11). Sin embargo, esto es sólo un tipo de sonrisa; hay muchas variaciones sobre ellas, teniendo diferencias de intensidad de actuación.

Una limitación bastante conocida es que carece de información temporal y espacial detallada[22], a escala local y global[23] [Pelachaud et al., 1994].

Existen estudios que abordan el asunto de si es posible o no una clasificación universal de expresiones faciales en emociones [Ekman & Friesen, 1971] dada la variabilidad fisionómica de cada persona observada, el género, la raza, y la edad. Además hay que tener en cuenta que cada emoción puede ser expresada en distintos niveles de intensidad, para ello, si los niveles de ambos son bastante bajos, las diferencias en su expresión facial son mínimas. En cambio, si su intensidad es mayor, las diferencias en

[22]No obstante fue diseñado para el estudio de imágenes estáticas

[23]Aproximaciones vistas en la extracción de características del rostro

el rostro son más fáciles de apreciar. Incluso la interpretación del lenguaje corporal (no sólo de la cara) es dependiente del contexto [Russell & Dols, 1997].

Finalmente, hay algunos estudios psicológicos que argumentan que el factor temporal de la expresiones faciales es un factor crítico en su interpretación [Bassili, 1978, Bruce & Voi, 1983, Izard, 1990]. Por esa razón, los investigadores se han decantado por sistemas de tiempo-real que analizan la cara como un todo, en un sistema dinámico que evoluciona con el tiempo, y en el que cada AU se puede inferir a partir de los AUs detectados previamente.

8.5.4. Trabajos completos que agrupan las tres fases

En 2006, Lucey [Lucey et al., 2006] se basa en la aproximación de modelos globales del rostro y adaptables a las observaciones (modelo holístico), así como en el sistema FACS. Para la clasificación, partiendo del modelo malla, se usan varios tipos de clasificadores: "vecino más cercano" y "*Support Vector Machine* (SVM)" [24]. Según los resultados experimentales reflejados no logra unas tasas de acierto muy elevadas en la detección de AUs (en media, entorno al 50-60 %).

En 2007, Yan Tong [Tong et al., 2007], propuso una aproximación que tiene en cuenta las relaciones internas entre cada AU y su evolución temporal, con el fin de mejorar la precisión y robustez. Para ello, se usa un red dinámica bayesiana para modelar las relaciones entre los distintos AUs. Con este sistema probabilístico se pueden deducir AUs no detectados por el sistema e visión artificial. Los experimentos muestran un incremento notorio en la precisión de reconocimiento de AUs, especialmente en casos reales.

Los métodos vistos hasta ahora, logran mejoras en la precisión gracias fundamentalmente al uso de métodos modernos de aprendizaje automático y a bases de datos con mayor número de muestras y calidad, pero no proporcionan el código fuente ni librerías que realmente permitan verificar los resultados expuestos. En un reciente trabajo (2011), Littlewort ([Littlewort et al., 2011]) presenta la herramienta visual **CERT**[25], que permite la clasificación de emociones en tiempo real (6 emociones), así como la detección de 19 FACS AUs.

Para la extracción de características usa la aproximación basada en puntos característicos (modelo local), y para la clasificación de cada posible AU, usa un algoritmo de SVM. La herramienta está disponible en: `http://mpt4u.com/AFECT`[26]. Sus creadores proclaman tasas de acierto promedio de entre el 80 % y el 90.1 % para el

[24] `http://en.wikipedia.org/wiki/Support_vector_machine`

[25] Entre los miembros que participan en su desarrollo, se encuentra como asesor Paul Ekman, con mas de tres décadas en el estudio de las emociones faciales y creador del comentado modelo FACS.

[26] Recientemente están trabajando en una versión comercial de dicha herramienta.

reconocimiento de emociones [27].

Otro interesante trabajo [Wierzbicki et al., 2012], es el desarrollado por *Fraunhofer Institute for Integrated Circuits in Erlangen* donde se presentan la librería **SHORE** (*Sophisticated High-speed Object Recognition Engine*). Una demostración funcional del software se puede descargar en: `http://www.iis.fraunhofer.de/en/bf/bsy/produkte/shore.html`. Esta librería es capaz detectar caras con robustez y velocidad [Küblbeck & Ernst, 2006], de reconocer expresiones faciales, identificar usuarios, detectar género (masculino o femenino), así como una edad aproximada. En el artículo ([Wierzbicki et al., 2012]) se comentan brevemente algunas de las aplicaciones reales, en especial las que tienen que ver con el ámbito del ocio y de la salud (soluciones asistenciales). Sin embargo, no existe literatura ningún dato acerca de la precisión en los reconocimientos automáticos.

Tal y como se acaba de ver, existen varios sistemas que proclaman tasas de acierto en el reconocimiento de las emociones a estudiar elevadas. Con dos de ellas, se ha probado intensivamente (ver la sección 8.7) y se han integrado dentro de nuestro sistema en forma de nodos ROS. Estos nodos ROS, forman el sistema que se ha denominado GEFA (Gender and Emotion Facial Analysis). La elección concreta de estas dos herramientas ha sido fruto de un estudio minucioso de la literatura y de las herramientas actuales disponibles.

8.6. Integración de GEVA y GEFA: sistema completo propuesto

El sistema final de detección de emociones es el fruto de la integración en paralelo de dos componentes como se ve a continuación (ver Fig. 8.5):

1. GEVA: el módulo de extracción de características sonoras y clasificación de emociones y género GEVA, desarrollado en lenguaje Chuck y explicado anteriormente. Este componente es capaz de ser ejecutado tanto en Ubuntu como en Mac[28]. Este módulo analiza la voz humana y proporciona como salida el género del usuario y uno de los cuatro posibles estados emocionales: neutro, feliz, triste, o sorprendido[29]. Internamente usa varios clasificadores para la emoción

[27] Sólo entre dos posibles alternativas: emoción positiva o negativa

[28] Pese a que es multiplataforma, existen varios factores que lo dificultan: implementación de Chuck, diferencia entre los sistemas de audio de ambos sistemas operativos: CoreAudio en MacOS y ALSA con portaudio en Ubuntu, el tipo de micrófono y tarjeta de sonido usados, el nivel de captura de audio usados en la fase de entrenamiento del mismo, etc.

[29] El valor de confianza en cada detección de emoción depende de la confianza asignada a priori en la fase de entrenamiento del clasificador

(y también para el género) con independencia del usuario y del idioma. El entrenamiento mediante la construcción del clasificador para el género y para el estado emocional se realiza *offline*. Sin embargo, el reconocimiento durante los experimentos con el sistema ya implementado se realiza en tiempo real (*online*).

2. GEFA: formado por:

   - El *third-party*[30] *SHORE* para el reconocimiento de expresiones faciales en el rostro. Este software proporciona como salida los valores de intensidad de los estados emocionales: felicidad, tristeza, sorpresa, y enfado (como se comento en el apartado 8.3.1, esta última no es tenida en cuenta). Además, si de estos posibles estados emocionales ninguno de ellos supera el 50 % de intensidad se supone estado neutro en el usuario (esto es necesario debido que no hay una emoción asociada al estado de neutro/tranquilidad). Por otro lado, también proporciona la edad con un margen de error variable de años, así como el género (sin valor de confianza en el mismo).
   - El *third-party CERT* también para el reconocimiento de expresiones faciales en el rostro mediante visión artificial. Este software proporciona los valores de intensidad de cada posible emoción. Las salidas son: diversión, contento, detector de sonrisa (estas tres se agrupan en un sólo conjunto como felicidad), disgusto y tristeza (estas dos se agrupan como tristeza), sorpresa, neutro, miedo, y enfado (estas dos últimas no son tenidas en cuenta, como se dijo en el apartado 8.3.1). Además del género.

La agrupación de emociones similares en una única emoción se realiza para que el conjunto de salidas de los distintos modos usados sean equivalentes. De esta manera es más sencillo establecer una regla que especifique cómo mezclar la información de salida de cada módulo y determinar la emoción predominante.

El objetivo final que se persigue es detectar la emoción predominante en el usuario e integrar esta salida del módulo de detección de emociones con el sistema de interacción RDS. Esta integración con RDS se realiza añadiendo, en cada acto comunicativo, que se corresponde con el turno de conversación, la información relativa al estado emocional del usuario (siempre y cuando sea posible su detección). De esta manera, en cada acto comunicativo están presente valores como los siguientes: la frase literal reconocida, los valores semánticos de la frase reconocida, la identificación del usuario, su posición espacial respecto al robot, la parte del cuerpo tocada, la pose del usuario, etc.

[30]Programa de terceras partes

8.6.1. Regla de decisión

En la Fig. 8.3 se ilustra el proceso de toma de decisión que determina la emoción predominante en el usuario. Se ha observado, y como tal se tiene en cuenta, que mientras que el usuario está hablando el sistema visual no determina correctamente la emoción. Por lo tanto, cuando el usuario esta callado sólo se tiene en cuenta la salida proporcionada por GEFA (visual), mientras que cuando está hablando se tiene en cuenta la salida de GEVA (sonoro). Una vez que la actividad de voz ha finalizado el acto comunicativo se considera que ha concluido. Teniendo en cuenta esta restricción temporal, es necesario definir una regla de decisión que combine la información dada por cada uno de estos dos módulos en una sola salida. La definición de esa regla depende de la tasa de acierto (precisión) de cada módulo por separado respecto a cada emoción a reconocer. Por ello, la regla de decisión es establecida después de realizar experimentos con GEVA y GEFA por separado.

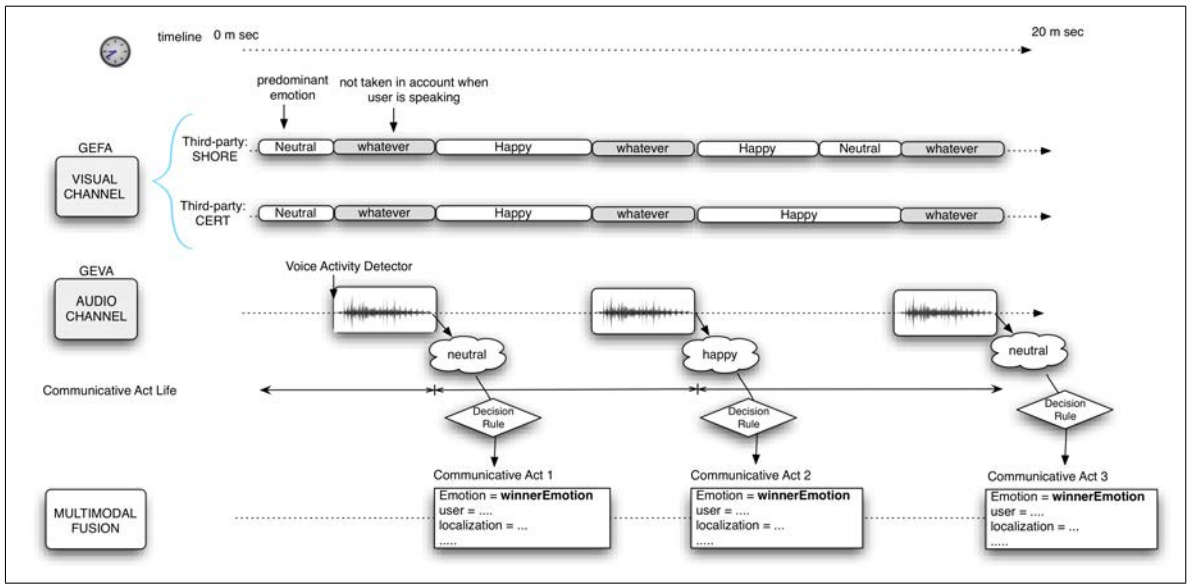


Figura 8.3: Esquema del proceso de toma de decisión para determinar la emoción del usuario en cada acto comunicativo

Una vez que se calcula la tasa de acierto de cada componente por separado (dos modos con dos clasificadores cada uno de ellos) y se construyen las matrices de confusión, es necesario establecer una regla que determina la emoción percibida predominante. Para ello, es necesario determinar el grado de certeza (confianza) que se tiene en la salida ofrecida por cada clasificador, ya que cada uno de ellos ofrece una posible emoción estimada sin determinar ningún grado de confianza en dicha estimación.

En este trabajo, el grado de confianza de cada salida es calculado usando el teorema de Bayes y las matrices de confusión previamente obtenidas. El teorema de Bayes es

aplicado de la siguiente manera:

Asumiendo que el estado actual es S y que la lista de estado emocionales es finita: $\mathbf{S} = \{s_1, \ldots, s_n\}$ donde $n \in \mathbb{R}$.

Entonces, se supone que existe un clasificador C, cuya salida es $S_C \in \mathbf{S}$, que se usa para estimar $S \in \mathbf{S}$.

La *matriz de confusión* $M \in \mathbb{N}^{n \times n}$ del clasificador C es definida como sigue:

$$\forall i, j \in [1, n], M_{ij} = p(S_C = s_j | S = s_i) \tag{8.1}$$

donde s_{C_j} es el estado estimado por el clasificador C.

Por lo tanto, el value en la fila i y columna j corresponde a la probabilidad que el clasificador estime el estado $S_C = s_j$ dado el estado real $S = s_i$.

Por ello, un clasificador ideal tendría una matriz de confusión diagonal tal que, $M_{ij} \neq 0$ si $i = j$ y $M_{ij} = 0$ si $i \neq j$. Esto es, que los estado detectados sea siempre igual que los reales.

Para determinar el grado de confianza del clasificador C, se necesita estimar la probabilidad de que el sistema de estar en un caso real s_i cuando la salida del clasificador es s_j. En otras palabras, se quiere calcular $p(S = s_i | S_C = s_j)$.

$$\begin{aligned}
p(S = s_i | S_C = s_j) &\underset{\text{(Bayes' theorem)}}{=} p(S_C = s_j | S = s_i) \frac{p(S = s_i)}{p(S_C = s_j)} \\
&= \frac{p(S = s_i) p(S_C = s_j | S = s_i)}{\sum_{k \in [1,n]} p(S_C = s_j \cap S = s_k)} \\
&\underset{\text{(Kolmogorov def.)}}{=} \frac{p(S = s_i) p(S_C = s_j | S = s_i)}{\sum_{k \in [1,n]} p(S_C = s_j | S = s_k) p(S = s_k)} \\
&= \frac{p(S = s_i) M_{ij}}{\sum_{k \in [1,n]} M_{kj} p(S = s_k)}
\end{aligned} \tag{8.2}$$

En el caso donde todos los estados son equiprobables, la probabilidad de s_i dado s_j es:

$$p(S = s_i | S_C = s_j) = \frac{M_{ij}}{\sum_{k \in [1,n]} M_{kj}} \tag{8.3}$$

Se supone que se tiene una colección de m clasificadores $\mathbf{C} = C_1, \ldots, C_m, m \in \mathbb{R}$.

Se quiere estimar S, el estado actual del sistema. Se sabe la salida de cada clasificador $S_C \in \mathbf{S}, \forall C \in \mathbf{C}$.

Como se ve, para cada posible estado del sistema $s_i \in \mathbf{S}$ y para cada clasificador $C \in \mathbf{C}$, se puede estimar $p(S = s_i|S_C)$.

Regla de decisión:

Conociendo todas las probabilidades condicionales, un estado altamente probable y fácil de determinar es el estado $\hat{s}$ que tiene la mas alta probabilidad de entre todos los clasificadores:

$$\hat{s} \in \mathbf{S}, \hat{C} \in \mathbf{C} \mid \forall s_i \in \mathbf{S}, \forall C \in \mathbf{C}\ ,\ p(S = s_i|S_C) \leq p(S = \hat{s}|S_{\hat{C}}) \tag{8.4}$$

8.7. Experimentos

Esta sección describe los experimentos realizados para determinar la precisión de GEVA y GEFA por separado, y de todo el sistema en su conjunto aplicando la regla de decisión establecida. Por lo tanto, hay un conjunto de experimentos iniciales para determinar el grado de acierto de ambos módulos. Estos experimentos son necesario para establecer la regla de decisión que une ambas salidas en una sola, la emoción predominante que se detecta. La regla de decisión es establecida después de analizar los vídeos de los experimentos y de considerar la tasa de acierto para cada emoción en particular.

Una vez que el sistema completo es definido, se llevan a cabo un conjunto de pruebas para medir su rendimiento.

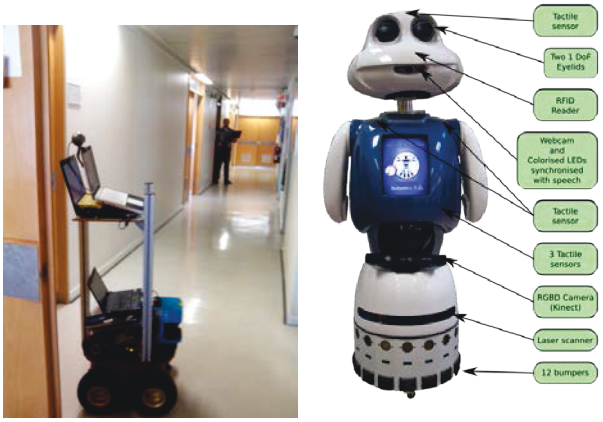
8.7.1. Configuración del experimento

La primera parte de los experimentos para medir el rendimiento de GEVA y GEVA, fueron hechos en varias sesiones en el centró robótico IST/ISR en Lisbon (Portugal) y en la Universidad Carlos III de Madrid (España). Antes de realizar experimentos con usuarios reales, se hicieron algunos test en el laboratorio para asegurarnos de la correcta configuración de los mismos.

Los experimentos se llevaron a cabo usando un robot Pioneer en Lisboa, y el robot Maggie en Madrid. (Fig. 8.4). En ambos, la configuración adaptada fue la misma, y es la que sigue:

- Un Apple Macbook con MacOS ejecutando CERT y GEVA. Este ordenador incorpora una cámara web *iSight* de 1.3 megapixeles y un micrófono omnidireccional con cancelación de ruido. La salida por pantalla y la entrada de audio, fueron almacenadas usando el programa Quicktime, para su posterior análisis.

- Un ordenador portátil con Windows 8 ejecutando SHORE con una cámara web externa de 2 megapixeles. La salida por pantalla también fue grabada.



(a) Robot Pioneer (b) Robot Maggie

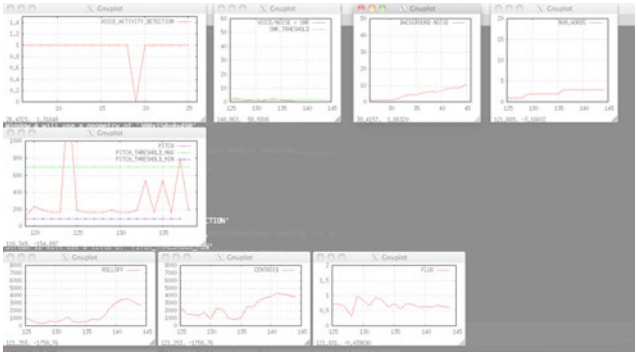
Figura 8.4: Los robots empleados en los experimentos

- Un netbook con Ubuntu 12.04 ejecutando un módulo de comunicación de sonidos no verbales [Alonso-Martin et al., 2012]. Este computador no es relevante en el análisis de los resultados experimentales, pero fue útil para captar la atención del usuario.

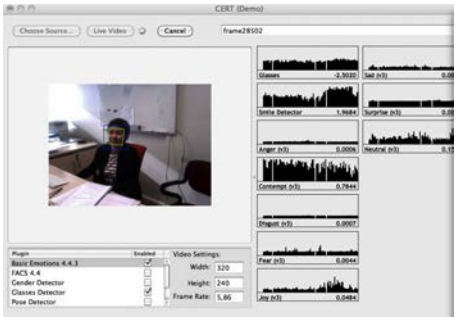
Unas capturas de pantalla del experimento realizado en Lisboa puede verse en la Fig. 8.5.

En los experimentos realizados en el IST/ISR, el robot primeramente se movía libremente por el entorno aproximándose a posibles usuarios. El robot trataba de llamar su atención, emitiendo sonidos no verbales (una especie de "lenguaje robótico"). El objetivo de esta primera parte del experimento era enganchar a los usuarios en una interacción con el robot y detectar cuales eran sus emociones surgidas de manera espontánea. En este caso, el robot fue teleoperado por los supervisores. El principal problema de esta aproximación fue que no todas las emociones posibles y contempladas por este trabajo fueron expresadas por los usuarios (básicamente, durante la interacción expresaron neutralidad y felicidad). Por ello, se modificó el experimento.

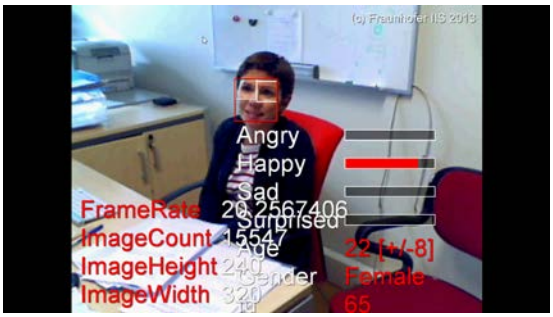
Un nuevo experimento se llevó a cabo con 40 estudiantes de grado de diferentes nacionalidades. En este caso, el supervisor del experimento pedía a los usuarios que simulasen las emociones que les eran requeridas. El experimento fue repetido durante un seminario para alumnos de doctorado, en una clase de 20 alumnos. De nuevo, el sistema fue probado con emociones simuladas. Durante este proceso de "interacción", la distancia entre el robot y el usuario se mantuvo por debajo de los 2 metros.



(a) GEVA



(b) CERT



(c) SHORE

Figura 8.5: Imagen tomada durante los experimentos llevados a cabo en el IST de Lisboa

Se realizó un nuevo experimento, ya en Madrid, para estudiar la relación entre dos factores concretos relacionados con GEFA: la distancia de interacción y la resolución de la cámara usada. Por ello, se remplazo la anterior cámara web de 2 megapixeles, por una *Logitech c920* de 15 megapixeles y óptica *Carl Zeiss*. Se realizaron algunas pruebas usando diferentes distancias de interacción.

Como ya se ha dicho, el primer conjunto de experimentos se hicieron para calcular la tasa de acierto de cada módulo, GEVA y GEVA, para cada posible emoción. Usando estos resultados, se fijó la regla de decisión que une ambas salidas. Con el sistema completo se han llevado a cabo nuevos experimentos en Madrid con 16 usuarios.

8.7.2. Experimentos con GEVA: audio

Los resultados obtenidos con este módulo fueron calculados usando frases, de al menos, 3 segundos de duración. Se han obtenido dos matrices de confusión, una por cada clasificador implementado (J48 y JRIP). Las filas muestran las emociones reales expresadas por los usuarios (espontáneas o simuladas) y las columnas muestran la emoción estimada por ese clasificador. La suma de toda la fila debe ser cien:

J48 classifier

	Feliz	Neutral	Triste	Sorpresa
Feliz	50	50	0	0
Neutral	0	80.28	19.71	0
Triste	0	33.33	66.66	0
Sorpresa	28.57	42.85	0	28.57

JRIP classifier

	Feliz	Nuetral	Triste	Sorpresa
Feliz	30	70	0	0
Neutral	0	87.32	12.67	0
Triste	0	22.22	77.77	0
Sorpresa	16.66	50	0	33.33

Cuadro 8.1: Matrices de confusión para GEVA

Si se mira las matrices de la tabla 8.1 se puede observar que el clasificador J48 no resulta muy preciso para distinguir el tono de voz feliz del neutro siendo el feliz el real. Sin embargo, sí resulta bastante preciso (80.28 %) para reconocer el estado neutro cuando así se expresa el usuario. Por otro lado, cuando el usuario esta triste es capaz de reconocerlo con una tasa de acierto del 66.66 %, además las veces que se equivoca determina únicamente que es neutro. Finalmente, no es muy preciso al tratar la expresión de sorpresa. En este último caso sólo acierta el 28.57 % de las veces, pero como dato positivo, se puede observar que nunca da un falso positivo de sorpresa[31]. Si se mira la matriz de la derecha, asociada al clasificador JRIP, resulta bastante preciso para detectar el estado neutral y triste, y muy poco preciso para reconocer el estado de felicidad o sorpresa.

[31]cuando se reconoce sorpresa se hace con total certeza, puesto que siempre que da esa salida corresponde a ese estado expresado por el usuario

Se ha observado que la tasa de error aumenta considerablemente cuando el usuario trata de simular la emoción que le pide el supervisor y no lo realiza de una manera muy expresiva (no lo expresa con un alto grado de intensidad). Es en esos casos, en los que incluso el propio supervisor le cuesta reconocer la emoción expresada por el usuario, cuando el número de fallos se incrementa. Esto de alguna manera queda reflejado en las matrices de confusión, puesto que se puede observar que los fallos producidos suelen recaer en determinar una emoción bastante cerca a la realmente expresada. Por ejemplo una emoción de felicidad expresada tímidamente el sistema la puede confundir con neutral. A su vez una emoción de tristeza expresada con poca intensidad puede ser confundida con neutral. Lo mismo sucede cuando el usuario se expresa su estado neutro de una manera muy relajada, en ese caso puede ser confundida con tristeza. Finalmente, sorpresa parece no estar relacionado con el resto de emociones.

El sistema GEVA, pese a que ha sido entrenado con varios idiomas, no resulta preciso con usuarios de varios idiomas. Como se apuntó en trabajos previos descritos en el estado del arte, la precisión del sistema se ve claramente disminuida cuando se trata de abarcar con un mismo clasificador todo tipo de usuarios, de todo tipo de idiomas. Para superar esta dificultad, sería conveniente tener un clasificador distinto para idioma e incluso dentro de ese idioma uno distinto para cada sexo. La idea de tener un clasificador individual para cada usuario no es bien recibida, debido a que haría falta para cada usuario una fase de entrenamiento con el sistema para aprender la voz en cada estado emocional[32]

8.7.3. Experimentos con GEFA: visión

Como se ha explicado, GEFA esta formado por dos software de terceras partes: SHORE y CERT. Las matrices de confusión para ambos son las siguientes:

Si se miran las matrices de confusión de la tabla 8.2, se puede observar en la de la izquierda asociada a SHORE, que éste resulta muy preciso reconociendo las emociones de felicidad (100 %) y de sorpresa (80 %), en cambio resulta algo menos preciso para reconocer el estado de neutro (66 %) y de triste (55.55 %). Si se mira la tabla de la derecha, referente a la tasa de acierto de CERT, se puede observar que es muy preciso reconociendo el estado de felicidad (100 %) y neutral (85.71 %), sin embargo es muy poco preciso para reconocer el estado de tristeza (28.57 %) y de sorpresa (36.36 %).

[32]En el campo del reconocimiento de voz, ha evolucionado de forma similar. Hace una década era necesario de que el usuario personalizase/entrenase el reconocedor leyendo varios textos en voz alta. En los últimos años, gracias a tecnologías como la de *Google* o *Nuance*, es posible reconocer voz en cualquier idioma y por cualquier usuario. No obstante, se piensa que internamente cada idioma tiene su propio reconocedor de voz, siendo el programa que usa el reconocedor de voz del idioma que se va a usar

SHORE

	Feliz	Neutral	Triste	Sorpresa
Feliz	100	0	0	0
Neutral	0	66.66	16.66	16.66
Triste	0	22.22	55.55	22.22
Sorpresa	0	10	10	80

CERT

	Feliz	Neutral	Triste	Sorpresa
Feliz	100	0	0	0
Neutral	0	85.71	14.28	0
Triste	0	71.42	28.57	0
Sorpresa	9.09	54.54	0	36.36

Cuadro 8.2: Matrices de confusión para GEFA

Como se ha dicho en la sección 8.6.1, durante los experimentos se ha observado que las expresiones faciales reconocidas cuando el usuario habla no son precisas. Esto ocurre porque el sistema FACS fue creado para el estudio de imágenes estáticas y no en movimiento. Cuando el usuario habla, abre la boca y los software usados detectan que el usuario esta feliz o sorprendido. En este caso, el sistema fluctúa entre esas dos emociones posibles.

Cuando el usuario esta callado, SHORE detecta relativamente bien la felicidad y el estado neutral Sin embargo, tiene importantes problemas para detectar el estado de enfado (no tenido en cuenta) y el de tristeza. Por otro lado, CERT tiene problemas para detectar los estados de enfado, sorpresa y tristeza.

Como se ha comentado, se han complementado los experimentos de Lisboa con otro experimento llevado a cabo en Madrid, usando una cámara web de alta resolución (15 megapixeles) y la misma configuración que anteriormente. Se ha observado que los algoritmos de detección de cara y emociones de CERT trabajan bien hasta 2 metros de distancia entre la cámara y el usuario, en ese rango, la precisión del sistema no se ve afectado por la distancia de interacción. Sin embargo, si la distancia es mayor de dos metros, el sistema de detección de caras tiene dificultad para localizar al usuario. Esto mismo ocurre para ambas cámaras, la de baja y la de alta resolución.

En el caso de SHORE, la distancia máxima de interacción es de 4.5 metros. De nuevo, la tasa de acierto no se ve afectada en ese rango, si crece la distancia por encima de los 4.5 metros el resultado del reconocimiento no es confiable. Tampoco se aprecia una mejora por usar una cámara de alta o de baja resolución como las usadas.

8.7.4. Experimentos con el sistema completo

Como se dijo en la sección 8.6.1, una vez que se han obtenido las matrices de confusión, se puede aplicar la regla de decisión para determinar cual es la emoción predominante en el usuario. Para entender esta regla, se va a aplicarla al siguiente ejemplo: el usuario saluda al robot y este último intenta detectar la emoción del usuario. La emoción real el usuario es tristeza.

Siguiendo nuestra aproximación, los estados y los clasificadores son los siguientes:
$\mathbf{S} = \{happiness, sadness, neutral, surprise\}$

$\mathbf{C} = J48, JRIP, CERT, SHORE$
Se asume que el robot recibe las siguientes salidas de cada clasificador usado:

1. J48 (GEVA) → Neutral (emoción estimada)
2. JRIP (GEVA) → Triste (emoción estimada)
3. CERT (GEFA) → Neutral (emoción estimada)
4. SHORE (GEFA) → Neutral (emoción estimada)

Aplicando la regla de decisión definida (8.4) se obtienen los siguientes resultados[33]:

$$p(Neutral_{real}|Neutral_{J48}) = \frac{80.28}{50 + 80.28 + 33.33 + 42.85} = \mathbf{0.38} \quad (8.5)$$

$$p(Sad_{real}|Sad_{JRIP}) = \frac{77.77}{0 + 12.67 + 77.77 + 0} = \mathbf{0.82} \quad (8.6)$$

$$p(Neutral_{real}|Neutral_{CERT}) = \frac{85.71}{0 + 85.71 + 71.42 + 54.54} = \mathbf{0.40} \quad (8.7)$$

$$p(Neutral_{real}|Neutral_{SHORE}) = \frac{66.66}{0 + 66.66 + 22.22 + 10} = \mathbf{0.67} \quad (8.8)$$

De acuerdo con estos resultados[34] el clasificador con el valor mas alto de confianza en su estimación es el de JRIP que estima que la emoción del usuario es tristeza. Esto puede parecer contradictorio dado que los otros tres clasificadores estiman que la emoción real del usuario es la de neutral. Sin embargo, la regla de decisión establecida determina que la salida del clasificador con mayor confianza, es la que se debe escoger como la emoción predominante.

En este caso el clasificador ganador no tiene una confianza suficientemente alta, es decir, está por debajo de un determinado umbral que se puede fijar experimentalmente, por lo que se determina que la estimación realizada por el sistema completo no es suficientemente confiable. Si el umbral es demasiado bajo se corre el riesgo de obtener muchos falsos positivos, es decir estimar una emoción diferente de la que realmente el usuario a expresado. Por contra, si el umbral es muy alto, el sistema de diálogo frecuentemente carecerá de información sobre la emoción del usuario, pero cuando la reciba se podrá fiar de la estimación realizada.

Con esta regla de decisión se está seleccionando lo mejor de cada clasificador, de manera que se usa cada clasificador para lo que realmente es bueno ya que tiene una alta tasa de acierto a priori.

En el ejemplo anterior, se podría devaluar la salida de JRIP (ya que es el único cuya salida es *tristeza*), para mejorar la confianza en los otros, ya que comparten la misma salida emocional: *neutral*. En este último caso, dependiendo de como sea este proceso de incremento/decremento, se podría llegar a tener más confianza en el estado de *neutral* que ofrece SHORE que sobre el estado de *triste* que ofrece JRIP.

[33] En este experimento se asume que todas las emociones son equiprobables, dado que se le pide al usuario que simule cualquiera de ellas.

[34] Los resultados mostrados son el valor mas alto dado por cada clasificador.

8.7.5. Rendimiento del sistema completo calculado estadísticamente

Se han obtenido experimentalmente las matrices de confusión asociadas a cada clasificador (J48,JRIP,SHORE, y CERT). Además, se ha establecido la regla de decisión que fusiona la salida de cada una de ellos, y que se acaba de mostrar su aplicación mediante un ejemplo. Sin embargo, todavía no se ha presentado el rendimiento del sistema completo.

Siguiendo el mismo proceder que ejemplo que se acaba de mostrar, se puede generar un mayor número de ejemplos. La salida de cada clasificador debe estar de acuerdo con la distribución de probabilidad recogida en su matriz de confusión. Sobre estos ejemplos generados, se puede aplicar la regla de decisión establecida. Por ello, se puede calcular la matriz de confusión del sistema general de manera estadística, simplemente generando un número alto de ejemplos y aplicando la regla de decisión descrita.

Para obtener una matriz de confusión del sistema completo de manera estadística, se han generado 10.000 ejemplos de posibles salidas de cada clasificador por computador. A esas salidas generadas, se le ha aplicado la regla de decisión. Comprobando si la clasificación ha sido correcta o incorrecta, se pueden obtener los valores de la matriz de confusión. Los resultados se muestran en la Tabla 8.3.

	Felicidad	**Neutral**	**Tristeza**	**Sorpresa**
Felicidad	100	0	0	0
Neutral	0	46.77	41.46	11.76
Tristeza	0	1.635	96.66	1.701
Sorpresa	4.345	2.639	2.60	90.411

Cuadro 8.3: Matriz de confusión calculada estadísticamente (las filas representan las emociones reales, y las columnas las emociones estimadas

La tasa de acierto calculada para todo el sistema es de 83 %. Este valor confirma que la combinación del uso de la información visual y vocal con esta regla de decisión mejora la tasa de acierto de cada uno de los clasificadores por separado: 56.37 % para J48, 57.10 % para JRIP, 75.55 % para SHORE, y 62.66 % para CERT.

8.7.6. Rendimiento del sistema de detección de emociones multimodal calculado mediante experimentos con usuarios

El último experimento con usuarios reales fue realizado en la Universidad Carlos III de Madrid, usando el robot Maggie. El sistema completo de detección de emociones fue puesto a prueba trabajando conjuntamente con el resto del sistema de interacción aquí presentado. Un total de 16 usuarios interactuaron con el robot de la siguiente manera: el usuario entra en el laboratorio, el robot mediante su sistema de interacción es capaz de situarse a una distancia de aproximadamente 1.5 metros y de frente al usuario. El robot explica al usuario la metodología del experimento. Por ello, le solicita que exprese mediante gestos faciales y voz una a una las emociones que le vaya solicitando, insistiendo en que trate de expresarse de la manera natural, sin sobre actuar. Una vez que el usuario ha expresado la emoción requerida, el robot le comunica una nueva emoción a representar. Este proceso se repite con las cuatro posibles emociones a reconocer.

En este experimento, nuevamente el contenido del mensaje transmitido por voz por el usuario no es importante para el experimento, en cambio si lo es el tono, volumen, etc. con el que se expresa. Las emociones expresadas por los usuarios fueron simuladas, ya que recordar que ciertas emociones como la de sorpresa o tristeza son difíciles de conseguir en interacciones reales, por ello se ha intentado que las emociones fueran simuladas, pero con el mayor grado de naturalidad posible.

La sala donde se realizaron los experimentos estaba en silencio y únicamente se encontraba el robot y el usuario. El robot tiene la capacidad de mantenerse a una distancia de interacción adecuada y mirando frontalmente al usuario.

De nuevo, analizando la información grabada, la matriz de confusión es obtenida para probar la utilidad del sistema completo. La matriz obtenida se puede ver en la Tabla 8.4. Como se observa, los valores mas altos para cada fila están situados a lo largo de la diagonal de la matriz, lo que ratifica un buen rendimiento del sistema de detección. En este caso, se ha obtenido una tasa media de acierto del 77 %. Este valor es muy cercano al valor calculado estadísticamente (83 %) y mejor que la tasa de acierto de cada clasificador por separado, que recordar es: 56.37 % para J48, 57.10 % para JRIP, 75.55 % para SHORE, y 62.66 % para CERT.

	Felicidad	**Neutral**	**Tristeza**	**Sorpresa**
Felicidad	100	0	0	0
Neutral	6	67	27	0
Tristeza	15	18.33	66.66	0
Sorpresa	0	12	10	78

Cuadro 8.4: Matriz de confusión calculada experimentalmente (filas: emociones reales; columnas: emociones estimadas).

Estos resultados prueban que el uso de varios modos (visual y sonoro) para detectar la emoción del usuario mejora la tasa e acierto obtenida por cada uno de ellos por separado. Ademas, el sistema completo para ser suficientemente bueno como para detectar las emociones del usuario. De hecho, no parece presentar importantes problemas para reconocer ninguna de las cuatro emociones contempladas. Entrando en detalle, el sistema propuesto es muy bueno detectando el estado de felicidad, relativamente bueno para sorpresa, y finalmente regular para reconocer tristeza y neutral, aunque con valores aceptables.

8.8. Resumen

Se ha descrito un sistema multimodal de detección de emociones automática aplicado a la interacción humano-robot que se integra dentro del sistema general de interacción propuesto. En cada turno de conversación, se suministra al sistema de interacción la emoción expresada por el usuario, siempre y cuando el sistema la haya detectado con suficiente certeza. La integración de esta información emocional con el sistema de interacción sirve para aumentar la adaptación entre robot social y el usuario.

Además de presentar su integración, se ha descrito en profundidad el componente de detección de emociones multimodal. Por un lado, se ha desarrollado el sistema GEVA, que analiza la voz del usuario para determinar a que emoción pertenece de entre las cuatro consideradas: feliz, neutra, tristeza y sorpresa. Para ello, se extraen ciertas características de la voz, que son fruto del análisis temporal, en frecuencia (usando la transformada rápida de Fourier), y en tiempo-frecuencia (usando la transformada Wavelet). Con estas características, y mediante técnicas de aprendizaje automático supervisado, se han construido dos clasificadores (JRIP y J48). Además, GEVA contiene un necesario y sofisticado mecanismo de detección de comienzo y final de voz, que es capaz de diferenciarla del ruido.

Por otro lado, se ha desarrollado el sistema GEFA que detecta las emociones mediante análisis del rostro, para ello se han integrado dos herramientas software:

SHORE y CERT. Finalmente, se han realizado experimentos para calcular la tasa de acierto de cada uno de estos clasificadores, como consecuencia de ello se han obtenido la matrices de confusión de cada uno de ellos.

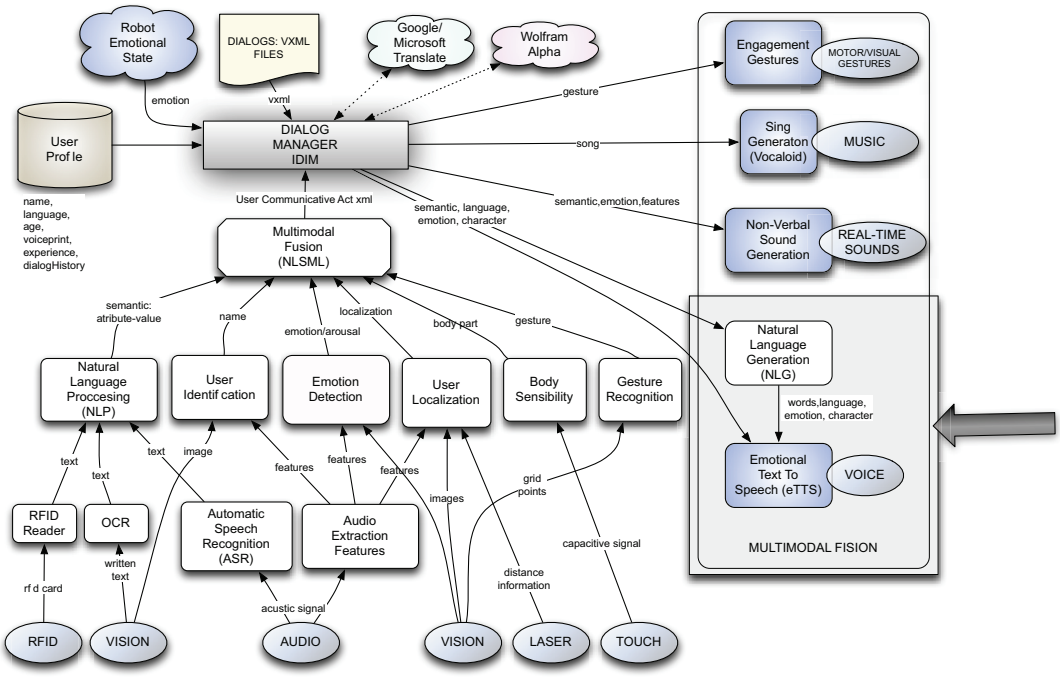
Con el uso de las matrices de confusión y utilizando la teoría Bayesiana se ha determinado una regla de decisión que logra fusionar las salidas emocionales de cada clasificador en una sola salida, que corresponde a la emoción predominante expresada por el usuario. Esta regla logra tener en cuenta las particularidades de cada clasificador para determinar la emoción correcta, de tal manera que si un clasificador es especialmente malo, la regla de decisión aplicada no lo tendrá en cuenta. Además, permite la inclusión de nuevos modos, con uno o varios clasificadores, sin necesidad de reformularse.

Finalmente, se han realizado nuevos experimentos sobre el sistema conjunto aplicando la regla de decisión establecida. Con ellos, se ha obtenido la matriz de confusión que indica la tasa de acierto de general de todo el sistema, que en promedio es de: 77 %. Estos resultados hay que tomarlos con cautela ya que la población sobre los que se han obtenido no es muy grande y las emociones expresadas por los usuarios han sido actuadas.

Como trabajo futuro queda la inclusión de nuevos canales en la detección de la emoción. Con la inclusión de nuevos modos, como son la información propia del contexto del diálogo, que la conoce el gestor del diálogo en base a su histórico; así como información semántica de la frase de voz reconocida, se puede mejorar la precisión en la interpretación que se hace de la emoción expresada por el usuario.

CAPÍTULO 9

Sistema de síntesis de voz con emociones



"Si buscas resultados distintos, no hagas siempre lo mismo."— Albert Einstein

9.1. Introducción

La tecnología que permite convertir texto escrito en una locución de voz se conoce como síntesis automática del habla. La síntesis de voz es, por lo tanto, la producción artificial de voz humana. El sistema informático usado para este propósito se conoce como sintetizador de voz, por lo que un sistema de "Text To Speech" (TTS), convierte texto escrito en voz. La síntesis de voz puede ser llevada a cabo como una concatenación de piezas de voz pregrabadas. Los sistemas difieren en el tamaño de las unidades de voz almacenadas; un sistema que almacena fonemas o difonos proporciona un mayor rango de salida, pero carece de claridad. Para dominios específicos, el almacenamiento de palabras enteras o oraciones permite alta calidad en la voz generada. Alternativamente, un sintetizador puede incorporar un modelo del tracto vocal, entre otras características de la voz humana, para crear una voz sintética de mayor calidad. En este trabajo, únicamente se trata de síntesis de voz, basada en la concatenación de fonemas y con un modelo del lenguaje, en ningún caso se usa un modelo que concatene palabras pregrabadas.

La calidad de la síntesis de voz es juzgada por su similitud a la voz humana y su capacidad de ser entendida con claridad por el humano. Un sistema adecuado de TTS permite a la gente con dificultades visuales o de lectura, escuchar textos escritos en un computador. En ese sentido, desde principios de los 90, muchos sistemas operativos y aplicaciones concretas han incluido este tipo de sistemas.

Un motor de síntesis de voz está normalmente compuesto por dos partes: un front-end (parte delantera) y un back-end (parte trasera). El front-end se encarga de dos importantes tareas. La primera, es convertir texto sin formato como números y abreviaciones en su equivalente en "palabras escritas". Este proceso es, a menudo, conocido como normalización del texto, pre-procesado, o "tokenización". Posteriormente, el front-end asigna transcripciones fonéticas a cada palabra, y divide el texto en unidades prosódicas: frases u oraciones. El proceso de asignar transcripciones fonéticas a palabras es conocido como conversión texto-a-fonema o grafema-a-fonema (en relación a transcribir la grafía a los sonidos que de el se deducen). La transcripción fonética conjuntamente con la información prosódica, componen la representación simbólica-lingüística de la salida del front-end. El back-end, se refiere a como el sintetizador convierte la representación simbólica-lingüística en sonidos. En algunos sistemas, esta parte incluye la computación de valores prosódicos como (el pitch, la velocidad, entonación, etc), que se desea aplicar a la salida del sistema.

Este es uno de los componentes clásicos de cualquier sistema de interacción ya que suele ser la manera más directa de transmitir información entre el sistema y el usuario. En sistemas de interacción como el que se está aquí presentando, aplicado al ámbito robótica social, este sistema pese a seguir siendo importante no es tan indispensable como en sistemas de diálogo telefónicos. De hecho, se pueden encontrar ejemplos de

robots, que únicamente se comunican mediante sonidos no verbales.

9.2. Requisitos

Las cualidades mas importantes que debe poseer un sistema que sintetiza voz son la naturalidad y la inteligibilidad. Por naturalidad, se entiende en como se parece el sonido sintético generado al generado por el humano, mientras inteligibilidad se refiere a la facilidad con la que la salida de voz es entendida. Un sistema de síntesis ideal debe poseer ambas cualidades, en ese sentido la mayoría de motores normalmente intentan maximizar ambas características.

El conjunto de los requisitos que se han establecido para nuestro sistema de TTS son los siguientes:

1. Voz claramente inteligible para los humanos.

2. La voz debe parecer robótica. Esto marca una gran diferencia con respecto a otros TTS, ya que no se pretende maximizar la naturalidad, sino la empatía con el robot. Como apreciaciones personales, en los experimentos llevados a cabo para este trabajo, parecen llevarnos a pensar que los usuarios muestran mayor empatía frente a las voces robóticas (pero con el mismo grado de inteligibilidad) que frente a voces humanas. Como se verá posteriormente se tienen varios sintetizadores de voz, algunos con voces humanas y otros con voces robóticas.

3. Capacidad de expresión con emociones , es decir, un tono de voz diferente para cada una de las emociones posibles: felicidad, tranquilidad, tristeza y nerviosismo. Debido a esta capacidad especial, en este trabajo se refiere al sistema aquí propuesto como (eTTS).

4. Capacidad de expresión en varios idiomas. Al enmarcarse dentro de un sistema de interacción multilenguaje es necesario que el TTS pueda expresarse en varios idiomas.

5. Sistema de control de la síntesis de voz centralizado en un único módulo, que da servicio a cualquiera de los otros componentes del sistema para el uso de la síntesis de voz. Dentro de este control de la síntesis de voz se deben encontrar funcionalidades como: gestión de la cola de locuciones, gestión del idioma y motor usado para sintetizar cada cola, servicios de traducción automática de frases, capacidad de interrumpir/pausar la síntesis en curso, control de volumen, etc.

6. Capacidad de expresar sonidos no verbales que ayuden a la interacción como: risas, bostezos, estornudos, suspiros, etc.

7. Múltiples "personajes de voz". Dado que el sistema puede funcionar en varios robots simultáneamente se desea que cada robot goce de su propio timbre de voz que lo diferencie del resto.

9.3. Principales sistemas de síntesis de voz

A continuación se describen los principales sistemas de síntesis de voz que se pueden encontrar actualmente:

- **Mbrola**[1]: Es un sistema de síntesis de código abierto que permite controlar la síntesis desde un nivel de abstracción bastante bajo, ya que permite un gran control prosódico, sin embargo, la inteligibilidad muchas veces no es la deseada.
- **Loquendo TTS** [2]: El software de síntesis de voz de la empresa Loquendo (actualmente forma parte de Nuance), sintetiza voz con un alto nivel de inteligibilidad en numerosos idiomas. Es el motor de síntesis preferido para la interacción desarrollada en esta tesis doctoral, ya que cumple todos los requisitos planteados previamente. Además posee gran poder expresivo, enfatizando palabras, añadiendo bostezos, risas, etc. Este sistema software ha sido famoso al poner voz a numerosos vídeos en youtube.
- **Nuance Real Speak**: Sistema de síntesis de voz de la empresa Nuance[3], que permite sintetizar voz en varios idiomas. Sin embargo, la calidad de la voz generada es menor que la de Loquendo TTS (ver ej. `http://www.nextup.com/nuance.html`.
- **Festival** [4]: Software libre de síntesis de voz. Con buena calidad para el idioma inglés, pero no tanto para el resto de idiomas.
- **Ivona**[5]: Sistema de síntesis de voz usado fundamentalmente en dispositivos Android y en productos de la empresa Amazon (recientemente ha sido adquirida por Amazon). La voz es de un alto nivel de inteligibilidad (muy similar a Loquendo TTS).
- **Pico TTS**: Voz por defecto en los sistemas Android (al menos hasta la versión más reciente, 4.2). Cumple con su cometido, pero la voz sintetizada se convierte en monótona. Desarrollado por la empresa SVOX.

[1] `http://en.wikipedia.org/wiki/MBROLA`
[2] urlhttp://www.loquendo.com/es/productos/sintetizador-de-voz/
[3] `http://www.nuance.com/`
[4] `http://www.cstr.ed.ac.uk/projects/festival/`
[5] `http://www.ivona.com/en/`

- **Google TTS**: Sistema de síntesis de voz usada por Google para todas sus servicios web. Normalmente corresponde a una mejora de las voces ofrecidas por Pico TTS. Se trata de una voz muy monótona y sin apenas variaciones ni en el tono ni en el ritmo.
- **Microsoft Reader**[6] Sistema de síntesis de voz usado por Microsoft para multitud de sus servicios web y aplicaciones. Se trata de un sistema de síntesis con un menor nivel de inteligibilidad que los anteriormente comentados. La voz suena demasiado monótona y robótica.
- **Verbio TTS** [7]: Sistema de síntesis desarrollado por la empresa catalana Verbio. La inteligibilidad de la voz no es tan alta como en Ivona o Loquendo, si bien está bastante cerca (probar la demo on-line: `http://www.verbio.com/webverbio3/index.php/es/demo-separator/demo-tts-online.html`. La calidad de la voz en español es superior a la inglesa.

En un análisis de sistemas de síntesis de TTS para la plataforma **Android**, descrito en la web: `http://www.geoffsimons.com/2012/06/7-best-android-text-to-speech-engines.html`, se puede encontrar la información resumida en la siguiente tabla:

Nombre	Libre	Calidad	Ritmo	Varianza	**Global**
Ivonna	Si	10	8	8	33
Svox Clásico (evolución de Pico TTS)	No	7	8	7	32
Loquendo TTS	No	7	6	10	27
Google TTS	Si	5	6	5	22
Pico TTS	Si	5	7	5	21

Cuadro 9.1: Resumen comparativo de diferentes motores de síntesis de voz

Los parámetros tenidos en cuenta han sido si es de pago o libre, la calidad del sonido[8], el ritmo del discurso[9] y la varianza[10]. La puntuación dada a cada uno de

[6] `http://en.wikipedia.org/wiki/Microsoft_text-to-speech_voices`

[7] `http://www.verbio.com/webverbio3/es/tecnologia/verbio-tts.html`

[8] ¿tiene suficiente resolución como para ser creíble?

[9] ¿el ritmo del discurso es coherente con lo que se podía esperar?

[10] Como el sistema es capaz de variar la manera en que enfatiza una palabra en función del contexto y de su función gramatical. Con eso se logra locuciones menos monótonas

ellos corresponde a una opinión subjetiva del autor, con la que como autor de este documento estoy bastante de acuerdo.

9.4. Descripción del sistema software

Los componentes de este módulo (ver Fig. 9.1) son: el *Wrapper/API*, permite la comunicación de los distintos módulos que quieren sintetizar voz con la habilidad de síntesis de una manera distribuida por la red; la *habilidad propiamente dicha de síntesis de voz*, se encarga de gestionar dichas peticiones de síntesis de voz, para sintetizarlas el momento oportuno, en el idioma adecuado, con el motor de síntesis oportuno, con el timbre adecuado al robot y con la emoción fijada; finalmente, en el nivel más bajo, se encuentran las *primitivas de síntesis* que implementan la síntesis de voz mediante distintos motores o herramientas libres/comerciales.

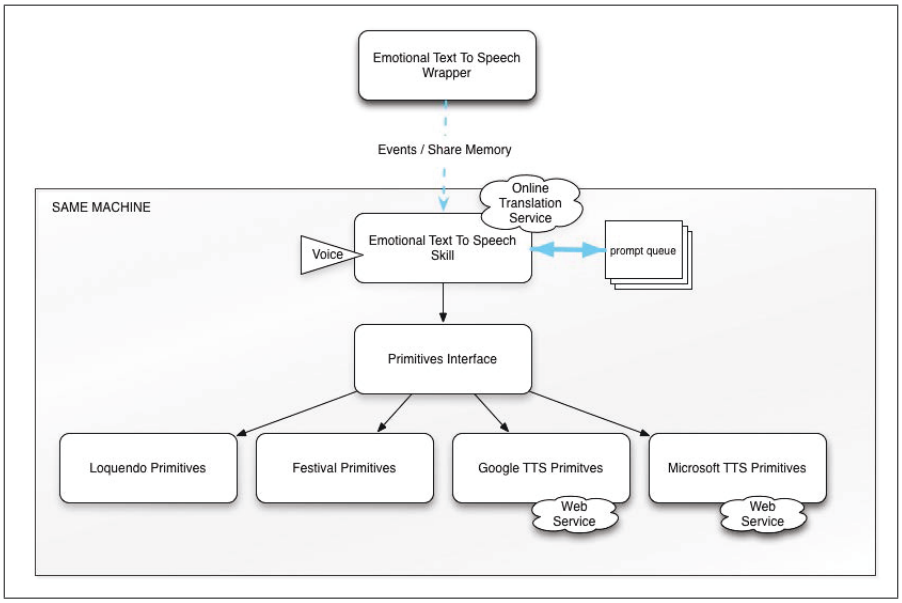


Figura 9.1: Arquitectura del sistema de síntesis de voz

El componente que actúa como Wrapper/API de la habilidad abstrae al resto de módulos de los mecanismos de comunicación necesarios para comunicarse con la habilidad que sintetiza voz. En este componente se ofrece toda la funcionalidad implementada en la habilidad. A continuación se enumeran las funcionalidades ofrecidas por el sistema:

- SetLanguage / GetLanguage
- SetEmotion / GetEmotion

- SetPreferedPrimitive / GetPreferedPrimitive
- SayText: Sintetiza el texto con la emoción, el lenguage y el motor establecidos previamente.
- SayTextWithEmotion: Sintetiza el texto con la emoción recivida.
- . SayTextNL: Permite el multilinguismo mediante traducción automática
- SayRandomSnippet: Permite cierto Natural Language Generation mediante plantillas
- SayTextGibberishLanguage: Permite comunicación en Gibberish Language1
- SayNonVerbalSound: Permite expresarse mediante sonidos en función del tag semántico, con un repertorio de sonidos previamente generados (en otro módulo se analiza la generación de sonidos pero en tiempo real).
- CanSpeakInmediatly: nos informa si se puede sintetizar en este mismo instante una nueva locución. Esto no siempre es así, puesto que puede haber locuciones encoladas o estar el sistema de voz secuestrado.
- IsSpeakingNow: Nos informa si el robot está actualmente hablando.
- ShutUp: Una vez que finaliza la síntesis de la locución en curso el robot se calla, se desechan todas las locuciones encoladas.
- ShutUpInmediatly: Similar a la anterior, pero deteniendo también la locución en curso.
- PauseSpeaking: Detiene la síntesis de voz en el instante actual, pero sin eliminar las locuciones encoladas.
- ResumeSpeaking: Continua la síntesis de voz en el mismo punto que se pauso mediante la función PauseSpeaking.
- ChanveVolume: Cambia el volumen al que se está sintetizando la voz.
- SetVolumeMax: Fija el volumen máximo.
- SetVoulmeMin: Fija el volumen mínimo.
- NumSpeechWaiting: Nos dice el número de locuciones encoladas esperando para ser sintetizadas.

- KidnapSpeech: Permite secuestrar la voz para que ningún otro módulo pueda usarla hasta que no se libere por el mismo.
- FreeSpeech: Libera el secuestro de la voz.
- IsVoiceKidnapped: Nos dice si el sistema de voz está secuestrado actualmente.

Como se puede ver por las funciones que ofrece el API, la habilidad de síntesis de voz debe implementar toda esta funcionalidad que permite trabajar con varios motores de síntesis, con emociones, con idiomas, con distintos timbres de voz y con sonidos pregenerados, todo ello de manera distribuida, coherente y concurrente temporalmente. La habilidad puede recibir peticiones de síntesis de manera simultánea, siendo necesario encolar en una estructura de cola dichas peticiones, para ser sintetizadas en el instante de tiempo oportuno. La habilidad de síntesis de voz actúa como un servidor capaz de ofrecer dicha funcionalidad teniendo el control de todo lo que ocurre en el manejo de la información y la síntesis verbal y sonora. Al estar en un sistema puramente distribuido se ha tenido especialmente cuiado en los mecanismos de sincronización, por semáforos, para evitar problemas de duplicidad o perdida de locuciones.

Finalmente la habilidad delega en las primitivas de voz la síntesis de las locuciones de voz. Esto permite que el sistema de síntesis de voz sea *plug-and-play*, en el sentido de que es muy sencillo añadir nuevos motores de síntesis de voz que mejoren/amplíen los existentes. Se ha implementado varias primitivas de voz basadas en varios motores: Loquendo, Festival, Google, Microsoft. Las dos últimas funcionan usando los servicios web correspondientes, que nos devuelven la locución en un fichero de audio que puede ser reproducido desde las primitiva. En nuestros robots sociales, la primitiva de voz usada por excelencia es la que usa el motor de síntesis de voz de Loquendo, ya que permite el uso de varios "agentes" que nos ha permitido trabajar con varios timbre y varias emociones simuladas. Además Loquendo nos ofrece un repertorio de "gestos verbales" como son risas, llantos, suspiros, silbidos, bostezos... que dotan de una notable expresividad a los sistemas de interacción.

Aunque no pertenezca propiamente al apartado de comunicación verbal, también se manejan primitivas de comunicación no verbal, por sonidos pregenerados y generados en tiempo real que deben ser gestionadas de la misma manera que las verbales por esta habilidad para evitar fenómenos de coarticulación verbal y no verbal.

Nuestro sistema utiliza por defecto las primitivas de voz basadas en Loquendo. Este motor de síntesis nos permite multitud de posibilidades, entre ellas la de modificar ciertos parámetros que generan una voz sintética mas "robótica" que "humana", control sobre ciertos parámetros prosódicos como es el volumen, velocidad de locución, frecuencia, cambios en la frecuencia, en el timbre, etc, que permiten expresar emociones con la voz. Todo ello, manteniendo la inteligibilidad de la voz generada.

Cabe resaltar un aspecto clave en cualquier sistema de diálogo moderno, y este es el multilinguísmo, en este sentido el reconocedor de voz (ASR), como ya se ha descrito, es capaz de trabajar con mas de una veintena de idiomas. El módulo de síntesis de voz, mediante las primitivas implementadas también es capaz de sintetizar en mas de una veintena de idiomas, así como hacer traducciones automáticas entre ellos. Para la traducción automática el sistema se ha valido de los servicios web de Google y de Microsoft que permiten dicha traducción online de frases. Basta con usar la función sayTextNL, de la siguiente manera:

SayTextNL ("es: Esto es una frase de prueba");

Si el idioma fijado previamente es el inglés, al decirle que sintetice el texto escrito en español, automáticamente debe ser traducido al inglés para su síntesis. La traducción NO se llevaría a cabo con la siguiente frase:

SayTextNL("en: This is a test sentence");

9.5. Generación de lenguaje natural

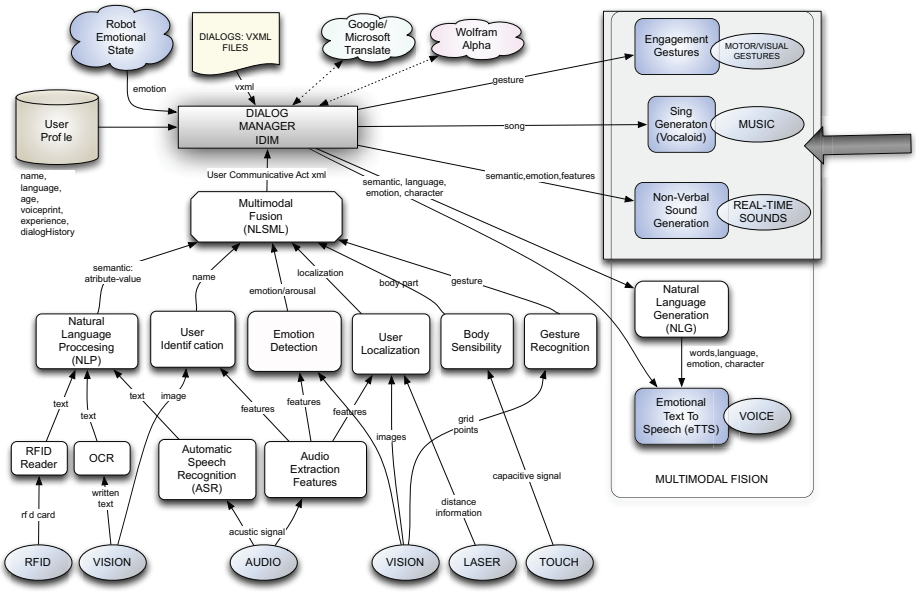
Se ha hecho un primer y básico acercamiento al Natural Language Generation (NLG). NLG posibilita la generación de frases partiendo de valores semánticos, con esto se consigue un cierta variedad en el repertorio a la hora de decir frases típicas como saludos, exclamaciones, despedidas, confirmaciones, evitando la reiteración constante de los mismos mensajes comunicativos. Esto se consigue gracias a la introducción de varias posibles frases asociadas a un valor semántico, de tal manera que cuando la habilidad intenta sintetizar un valor semántico este se traduce en tiempo real por uno de las posibles frases a sintetizar. Para la generación de sonidos no verbales se está siguiendo este mismo esquema, basado en valores semánticos y un posible conjunto de sonidos no verbales a generar en tiempo real.

9.6. Resumen

Se ha integrado dentro de la arquitectura de control un sistema de síntesis de voz. El sistema permite, como características mas relevantes: multilingüismo con traducciones automáticas, simular la expresión de emociones en la voz, varios timbres de voz asociados a cada robot, varios motores de síntesis de síntesis de voz, gestión de la cola de locuciones, generación de lenguaje natural mediante plantillas (en una versión muy básica). Este sistema de síntesis de voz se integra con el sistema de diálogo RDS, posibilitando que todos los componentes de la arquitectura de control pueden sintetizar voz.

CAPÍTULO 10

Otras modalidades de expresión



"Sueño con instrumentos obedientes a mi pensamiento y que con su aportación de un nuevo mundo de sonidos insospechados, se prestarán a las exigencias de mi ritmo interior"— Edgar Varese

10.1. Introducción

El sistema de interacción aquí presentado también permite al robot, cantar, bailar, y generar sonidos no verbales en tiempo real. La expresión de sonidos está relacionada con el análisis del ambiente sonoro, mediante mecanismos de extracción de características sonoras. Esto permite cierto tipo de adaptación al entorno sonoro, de tal manera que en función de dicho entorno se pueden generar respuestas sonoras adecuadas. El sistema, además es capaz de hacer que el robot actúe como un cantante, mediante canto y pasos de baile. Todas estas habilidades expresivas están integradas dentro del sistema de interacción multimodal aquí presentado.

Los robots sociales han sido creados y diseñados con el principal objetivo de interactuar con los humanos en un modo similar al que los propios humanos lo hacen entre ellos. Bailar y cantar, como arte de expresión humana, podría ser visto como una articulación de patrones que se pueden imitar y que se pueden encontrar en la interacción social. Sociólogos y antropólogos apoyan este punto de vista basándose en el análisis de la interacción social ([Grammer, 1998, Shaffer, 1982, Kendon, 1970]). Esos análisis muestran que la producción coordinada de movimientos y sonidos puede ser descrito como un baile.

En los sistemas de diálogos multimodales que manejan la interacción entre los usuarios y el robot, como en [Wahlster, 2003a][Niklfeld et al., 2001] [Shuyin et al., 2004] [Reithinger & Alexandersson, 2003] [Bohus et al., 2007], normalmente la multimodalidad es referida a los mecanismos de entrada al diálogo, y no tanto a las salidas expresivas del mismo. En un sistema de diálogo clásico, o incluso en uno moderno, las entradas y las salidas son normalmente limitadas a la combinación de voz y visión. Sin embargo, nuestro sistema implementado en el robot social Maggie (entre otros), es mucho más ambicioso ya que no sólo las entradas son multimodales sino que las salidas también lo son. Por lo tanto, el robot es capaz de expresarse por si mismo, mediante sonidos, melodías cantadas, música, voces emotivas y gestos.

El diálogo que maneja la interacción con el robot social, permite su control y la gestión de las salidas y entradas multimodales de una manera sencilla. Como ya se ha comentado en capítulos anteriores, para esta gestión de las entradas y salidas, es necesario escribir el o los diálogos en ficheros basados en una extensión del lenguaje Voice XML [1]. Partiendo de este lenguaje y su extensión, podemos representar la generación musical, de gestos de baile o de interacción en general. Por lo que, solamente es necesario especificar, en tales ficheros xml, cuales y cuando son generados/sintetizados.

[1] http://www.w3.org/TR/voicexml21/

10.2. Los lenguajes musicales

Hay que recordar que todos estos logros en la generación musical mediante ordenador no se podrían haber alcanzado sin la creación de Max Mathews y su lenguaje MUSIC en 1957. Es hora de que hablemos brevemente de la evolución y diferentes vertientes que surgieron a partir de éste. Aunque diferentes personas y centros de investigación colaboraron y crearon diferentes lenguajes, las características comunes de todos éstos hacen que se engloben dentro de una misma familia denominada "*MUSIC N*". La creación del primer MUSIC impulsó las investigaciones en distintos centros de Estados Unidos y Europa. Las diferentes versiones surgidas se identificaron por los números, no siempre progresivos, que iba tomando la N. La razón de no ser siempre progresivos tiene que ver con la creación de un lenguaje completamente nuevo y no sólo una mejora a partir de versiones anteriores. Music en su versión de 1986 acabó denominándose *CSound*.

Las características comunes que todas estas versiones compartieron en sus inicios fueron el enfoque alfanumérico a la hora de definir los parámetros y características musicales, el uso de Unidades Generadores y la utilización de ellos en tiempo diferido.Las Unidades Generadoras nacieron en el MUSIC III creado también por Max Mathews y se tratan probablemente de la mayor evolución en este tipo de lenguajes. En aquellos años era un enfoque revolucionario y altamente innovador; cuya importancia se manifiesta por el uso constante a lo largo de los años hasta la actualidad. Se trata de macros que realizan distintas funciones útiles tanto para la generación y el control de los sonidos. Las Unidades Generadoras son tanto los osciladores, filtros, envolventes de la amplitud como el retardo, la especialización de los sonidos y un largo etcétera. En cuanto al tiempo diferido, se trataba claramente de una característica no deseada, pero se encontraba ligada a los límites impuestos por el hardware de la época. No es hasta el nacimiento de *CSound* cuando un lenguaje musical permite operar en tiempo real.

Cmix, *CMusic* y *CSound* son lenguajes utilizados en la actualidad. De esta familia de lenguajes han nacido el resto de lenguajes actuales, tanto con Unidades Generadores gráficas como alfanuméricas. Los más utilizados, por mencionar algunos son: *Max/Msp*, *AudiMulch*, *SuperCollider*, *Pure Data*, *Jsyn*, *Common Lisp Music* y *Chuck*.

En las tablas: Tabla10.1 y Tabla 10.2, se ha tratado de resumir las características mas importantes de todos ellos.

Name	Purpose	First release date	Most recent version	Cost	License	User interface
Audulus	Realtime synthesis, audio processing, mobile music	2011/09	v1.9	$14.99 (iPad), $39.99 (Mac)	Proprietary	Graphical
ChucK	Realtime synthesis, live coding, pedagogy, acoustic research, algorithmic composition	2004	v1.3.1.2	Free	GPL	Document
Csound	Realtime performance, sound synthesis, algorithmic composition, acoustic research	1986	v5.15	Free	LGPL	Document, graphical for realtime
Impromptu	Live coding, algorithmic composition, hardware control, realtime synthesis, 2d/3d graphics programming	2006	v2.5	Free	Proprietary	Document
Max / MSP	Realtime audio + video synthesis, hardware control	1980s (mid)	v6.0.1	$400.00	Proprietary	Graphical
Nsound	Realtime synthesis, offline audio rendering, algorithmic composition, acoustic research	2003	0.8.1	Free	GPL	Document
Pure Data	Realtime synthesis, hardware control, acoustic research	1990s	v0.43	Free	BSD-like	Graphical
Reaktor	Realtime synthesis, hardware control, GUI design	1996	5.6.1	$399	Proprietary	Graphical

Sfront	MPEG-4/SA implementation, Realtime synthesis,algorithmic composition, structured Web audio	1997	0.96	Free	BSD-like	Document
Super Collider	Realtime synthesis, live coding, algorithmic composition,acoustic research, all-purpose programming language	1996, March	v3.5	Free	GPL	Document
Usine	Audio manipulation, live coding, algorithmic composition	2006	v5.20	Free o 120 euros	Proprietary	Graphical

Cuadro 10.1: Comparación de los principales lenguajes de programación musical

Name	Operating system(s)	Source code language	Programming (plugin) API language(s)	Other technical features
Audulus	IOS	Mac OS X	C++, Objective-C, Lua	
ChucK	Mac OS X, Linux, Windows	C++		Unified timing mechanism (no separation between audio-rate and control-rate),command-line access
Common Music	Mac OS X, Linux, Windows	Scheme, C++	Scheme, SAL	Command-line access
Csound	Mac OS X, Linux, Windows	C, C++	C; also Python, Java, Lisp, Lua,Tcl, C++	IDE (QuteCsound), multi-track interface (blue); several analysis/resynthesis facilities; can compute double-precision audio; Python algorithmic composition library
Impromptu	Mac OS X	Lisp, Objective-C, Scheme	C, C++, Objective-C, Scheme	Native access to most OS X APIs including Core Image, Quartz, QuickTime and OpenGL. Impromptu also includes its own statically typed (inferencing) systems language for heavy numeric processing - OpenGL, RT AudioDSP etc..
Max / MSP	Mac OS X, Windows	C, Objective-C	C, Java, JavaScript, also Pythonand Ruby via externals	
nSound	Mac OS X, Linux, Windows	C++	C++, Python	Real-Time Dynamic Digital Filters

Pure Data	Mac OS X, Linux, Windows, iOS, Android	C	C, C++, FAUST, Haskell, Java,Lua, Python, Q, Ruby, Scheme, others	
Reaktor	Mac OS X,Windows			
SuperCollider	Mac OS X, Linux, Windows, FreeBSD	C, C++, Objective-C	C++	Client-server architecture; client and server can be used independently, command-line access
sfront	Linux, Windows(via cygwin)	C++		Conforming MPEG-4/SA implementation
Usine	Windows	Delphi	C++	

Cuadro 10.2: Detalles técnicos de los lenguajes de programación musical

10.3. Implementación de la expresión de sonidos no verbales y canto

Recordemos al robot R2D2[2], de la famosa saga de la *Guerra de las Galaxias*, o al robot *Wall-E*[3], en la película homónima; cuando se comunicaban con ellos, respondían mediante sonidos no verbales, que acompañado de movimientos y gestos conseguían transmitir un mensaje comunicativo. Dado que se ha querido dotar al sistema de la capacidad de generar sonidos no verbales en tiempo real, se ha usado el lenguaje de programación musical Chuck, el mismo que se ha usado para la extracción de características del entorno sonoro. Este lenguaje permite fácilmente conectar el análisis con la síntesis del sonido.

La generación de sonidos, en primera instancia, trata de imitar o realizar un "eco sonoro" de la voz recibida por parte del usuario o de los sonidos que recibe de su

[2]http://en.wikipedia.org/wiki/R2-D2
[3]http://www.disney.es/wall-e/

entorno. Este eco, se realiza en el "lenguaje sonoro" del propio robot. Se sintetiza en tiempo real y constituye una manera "simpática" de responder el robot al usuario. Normalmente este tipo de expresión constituye sólo un complemento al resto de mecanismos expresivos: voz y gestos fundamentalmente.

Además de los sonidos sintetizados en tiempo real, existen sonidos pregenerados que permiten desde el propio diálogo expresar ciertas emociones o estados del diálogo, como son: error, emergencia, pensando, sorpresa, saludo, diálogo, etc. Por cada uno de estos "valores semánticos" existen varios sonidos pregrabados que se activan desde la especificación del diálogo en el XML correspondiente.

La generación en tiempo real de sonidos, permite realizar además acompañamientos musicales, puesto que el sistema de extracción de características sonoras es capaz de determinar la tonalidad de una canción o de una melodía de voz. Mediante el propio lenguaje Chuck se pueden sintetizar sonidos que se ajusten perfectamente a dicha tonalidad sonora.

En otro nivel, se ha dotado al sistema de la posibilidad de expresarse musicalmente mediante voz cantada. Es decir, el sistema permite que el robot cante ciertas canciones. Estas canciones están tabuladas mediante un programa de expresión musical llamado *Vocaloid III* (ver Fig. 10.1). Este programa permite sintetizar voz en tiempo real, cantando la letra con una determinada melodía. Para ello, sobre una partitura se escribe cada vocal a generar en una determinada altura (frecuencia) y con determinados efectos sonoros. Vocaloid ha sido desarrollado por Yamaha Corporation con la colaboración de la universidad Pompeu Fabra de España. La primera versión de este software sólo permitía el lenguaje Japonés, en la siguiente versión se incorporó el inglés, y finalmente en la tercera versión se incorporó, entre otros, el lenguaje español.

10.4. Gestos expresivos

Los seres humanos, cuando nos se comunican con otros humanos, a parte del uso de la voz, se suelen apoyar en gestos que complementan el significado de lo transmitido verbalmente. De esta misma forma, el sistema de interacción necesita de gestos que sirvan para aumentar la expresividad del robot. Por ejemplo, es deseable que si el robot detecta un mal entendido en la interacción, nos lo haga saber mediante el uso de la voz, al mismo tiempo que realiza un movimiento de izquierda a derecha de su cabeza girando el cuello, así como generando un sonido no verbal que indique frustración (como los que se han descrito en la sección anterior).

El sistema de gestos desarrollado para el sistema de interacción general, se basa en un *gestionario*, que es un conjunto numerado de gestos como saludos, asentimientos, negaciones, alzamiento de brazos, guiños, etc. Algunos de estos gestos basan su movimiento en la lectura de la información sensorial, por ejemplo, el gesto que posibilita

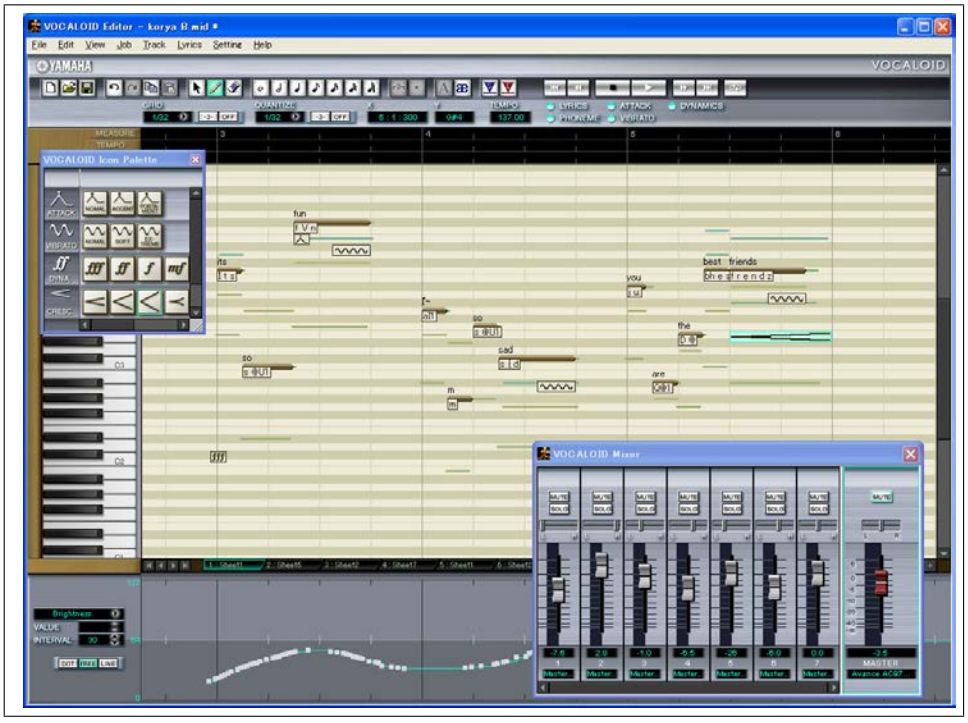


Figura 10.1: Interfaz para construir canciones sintéticas mediante Vocaloid

(a) Robot performing a gesture: Yupi

(b) Yupi gesture too

(c) Another Robot gesture: High C

(d) Robot gesture: Arrogant

Figura 10.2: Robot social Maggie realizando gestos de enganche para el diálogo

al robot seguir con la mirada al usuario mediante movimiento de su cabeza, mientras que otros gestos pueden ser configurados como un bucle infinito (ver Fig. 10.2). En cual quier caso, todos los gestos se han programado siguiendo una evolución temporal determinada y pueden ser activados/desactivados desde la especificación XML del propio diálogo. La activación de cada gesto puede ser realizada inmediatamente o quedando encolada la petición de activación hasta que finalice el gesto actual.

10.5. Resumen

Se han descrito las capacidades de expresión del sistema de interacción que permite al robot social Maggie comunicarse mediante gestos, sonidos no verbales y melodías de voz. Además, los diversos modos de expresión tienen en cuenta las entradas sensoriales del sistema, para adecuarse a ellas. En este sentido, la extracción de características sonoras juega un importante papel en la síntesis de sonidos.

Todas estas capacidades de expresión multimodal han sido integradas en el sistema de interacción de tal manera, que el gestor de diálogo es capaz de interpretar y ejecutar todos estos modos. Con ello se incrementa la naturalidad y versatilidad de la interacción producida.

Los sistemas expresivos pueden ser mejorados. La generación de sonidos no verbales, actualmente se basa en una política de "imitación", esto es, el sistema genera sonidos cuyos parámetros dependen directamente de los parámetros percibidos en la voz captada del usuario. Sin embargo otras políticas pueden ser deseables. Por ejemplo, contrapunto, o imitación con variación, etc.

Se están desarrollando nuevas teorías sobre expresión/síntesis automática de sonidos en un nivel morfológico ("sonemas"), que describen el sonido como una síntesis granular parametrizada, al igual que los sonidos verbales describen la oración verbal como un conjunto de fonemas regidos por ciertas reglas. Sin embargo, sería necesario desarrollar un modelo que relacione los parámetros morfológicos con alguna representación semántica de los mismos, con la intención comunicativa del robot o con la emoción que se quiere expresar.

El sistema gestual desarrollado tiene limitaciones en las que se está trabajando. Una de ellas es la falta de sincronización entre los gestos y el resto de modos expresivos. En este sentido, al igual que existe un módulo encargado de realizar la fusión multimodal, sería necesario un mecanismo que controlase la fisión multimodal en actos comunicativos de expresión, que tuviera en cuenta la evolución temporal de los distintos modos de expresión.

Otra limitación importante, es que los actuales gestos, se han desarrollado "a medida" para el robot Maggie, lo cual dificulta su incorporación a los nuevos robots que se están desarrollando. En este sentido, se necesita trabajar para aumentar la generalidad del sistema de gestos.

Sería conveniente, para el caso de robots con importantes capacidades de expresión facial, incorporar el sistema de gestos *FACS* (descrito en el capítulo que describe el sistema de detección de emociones), al sistema de expresión de gestos.

CAPÍTULO 11

Conclusiones y trabajos futuros

"Nunca olvides que basta una persona o una idea para cambiar tu vida para siempre, ya sea para bien o para mal"— Brown, J.

11.1. Conclusiones

Una tesis doctoral consiste en un proceso de investigación que concluye con un informe en donde se recoge el planteamiento del problema/s dentro de un área científica. Se explica lo que se sabe de él previamente, se describe el proceso llevado a cabo para resolverlo, se exponen los resultados obtenidos. Finalmente, se proponen futuras líneas de investigación para mejorarlo.

Partiendo de esta definición de tesis doctoral, se considera que el trabajo aquí presentado se ajusta a ella. La presente tesis se engloba dentro del campo de investigación que trata la interacción humano-robot. El problema o desafío principal planteado consiste en mejorar los sistemas actuales de interacción entre humanos y robots, de tal forma que la interacción resultante sea parecida a la producida entre humanos. Para ello, se ha planteado un sistema de interacción general, Robotics Dialog System (RDS). La introducción de este sistema divide el problema principal en varios más pequeños. Para dar solución a cada uno de estos subproblemas se ha analizado el estado del arte, se ha descrito el proceso seguido para darle respuesta, y finalmente se presentan las soluciones alcanzadas. En este sentido, la tarea llevada a cabo, no sólo ha consistido en un proceso de investigación, sino que también ha sido necesario un importante esfuerzo en tareas de integración, para lograr un sistema robusto y estable en el el tiempo.

Las aportaciones generales que aporta el sistema RDS son las siguientes:

- Sistema de interacción general, llamado RDS, capaz de trabajar con múltiples modos de interacción tanto a la entrada del sistema (fusión de la entrada multimodal), como a la salida expresiva del mismo (fisión multimodal). Entre estos modos, destacan principalmente los relacionados con el audio (*sistema multisonido*), siendo capaz de utilizar la entrada sonora para las siguientes tareas: reconocimiento de voz, identificación del usuario, localización espacial del usuario respecto al robot, detección de emociones del usuario; mientras que los canales de salida se información se usan para: síntesis de voz con emociones, síntesis de sonidos no verbales y generación musical. Al modo sonoro (*multisonido*) se lo complementa con el resto de modos para completar la multimodalidad: sistema visual, gestual, de radio frecuencia, y táctil.

- La entrada multimodal del sistema se gestiona inspirándose en la *teoría de actos comunicativos* que tiene lugar en la interacción entre humanos. De esta manera, se abstrae al "gestor de la interacción" de los modos/canales mediante los que se obtiene el mensaje comunicativo. Cada *acto comunicativo* percibido, consiste en un paquete de información formado por pares atributo-valor relativos al diálogo multimodal. Se ha desarrollado el componente de fusión multimodal, capaz de

agrupar estos valores semánticos de manera coherente en el tiempo, en paquetes de información que se intercambian a cada turno de la interacción.

- El Sistema de interacción es adaptable al usuario. Esta adaptación se realiza mediante el uso de perfiles de usuario que el sistema es capaz de generar y actualizar mediante el propio sistema de interacción natural. Estos perfiles contienen información del usuario relativa a su propia naturaleza y a distintas preferencias, como se explica a continuación:

  1. Adaptación al idioma (multilingüismo): el sistema es capaz de hablar en el idioma con el que se expresa el usuario. Esta adaptación es válida para cualquier idioma reconocido, pero funciona especialmente bien para español e inglés. Basta con que un usuario conocido salude al sistema para detectar la identidad del usuario y el idioma usado en la comunicación.
  2. Adaptación de la distancia de interacción (proxémica): el sistema es capaz de determinar cual es la distancia de interacción a mantener con cada usuario o grupo de usuarios.
  3. Adaptación a la familiaridad con el sistema: dada la experiencia de uso del usuario con el sistema (en este caso el robot), la interacción puede ser adaptada. De este modo, un usuario novel con el sistema puede ser "tutorizado" por el propio robot en su uso, mientras que un usuario "experimentado" apenas necesita de diálogos de aclaración o de sugerencias.
  4. Adaptación emocional: el sistema puede adaptar la interacción a la emoción detectada en el usuario, así como al propio estado emocional del robot.

Vista las aportaciones generales, a continuación se describen las **aportaciones particulares** realizadas en cada componente que forman parte de RDS.

- Se ha desarrollado un sistema de reconocimiento de voz capaz de trabajar simultáneamente con dos modos diferentes y complementarios: uno sujeto a gramáticas semánticas o otro en modo reconocimiento de texto libre mediante un modelo estadístico del lenguaje (ver capítulo 6 y apéndice C). Para ello, se realiza el reconocimiento de la voz usando dos reconocedores concurrentemente. Por otra parte, se han analizado las posibles configuraciones software-hardware necesarias para realizar esta tarea de manera satisfactoria.

- Se ha desarrollado un sistema multimodal de localización de usuarios respecto al robot, que se ha integrado dentro del sistema de interacción propuesto. Con esto y un conjunto de reglas (obtenidas tras un estudio proxémico de interacciones entre usuarios y el robot Maggie), se logra adaptar la distancia de interacción a cada usuario concreto o grupo de usuarios [Alonso-Martín et al., 2012].

- Partiendo de una tesis anterior y del estándar VoiceXML, se ha integrado y extendido un gestor de diálogo basado en el rellenado de huecos de información. De este modo, se permite en la propia especificación del diálogo, el uso de funcionalidades no incluidas en el estándar VoiceXML y sí proporcionadas por nuestro sistema, como son: la multimodalidad de entrada y salida, el multilingüismo, consultas a web semánticas, uso de servicios de reproducción de música en-linea, etc.

- Se ha desarrollado un sistema de síntesis verbal y no verbal[1], capaz de sintetizar voces y sonidos "con emociones". El sistema desarrollado, logra sintetizar voz y sonidos usando diferentes tecnologías, algunas propias y otras de terceras partes. Tanto el sistema verbal como el sonoro es capaz de expresarse simulando emociones.

- Se ha desarrollado un sistema de detección de emociones multimodal integrado con el sistema RDS. El sistema se basa en el análisis de la voz y del rostro del usuario. Esta detección emocional se puede usar para adaptar los diálogos y evitar el fracaso en la interacción.

11.2. Trabajos futuros o en desarrollo

A continuación se enumeran y describen las mejoras al trabajo presentado, y en las que ya se está trabajando:

- **Conversación multiparte**: el sistema aquí presentado está desarrollado para interacciones full-duplex[2] uno a uno. Hasta la fecha, no es posible una interacción coherente y simultanea entre varios usuarios y el robot. Para lograr esta tarea se presentan varios retos a superar, entre los que cabe reseñar:

  - Separar en diferentes canales cada modo. Cada modo, deberá tener tantos canales como usuarios haya presentes en la comunicación. Por ejemplo, el reconocedor de voz deberá recibir la voz de cada usuario por separado.
  - Gestión avanzada del turno de comunicación. El sistema deberá determinar a que usuario/s desea dirigirse en cada turno.
  - Carga simultánea de varios perfiles de usuario. Se necesitará que el sistema pueda cargar todos los perfiles de los usuarios presentes en la conversación y alternar entre ellos según corresponda.

[1] Recordar al robot R2D2 de *La guerra de las galaxias*

[2] ambas partes pueden iniciar conversación, e interrumpir el turno en cualquier momento

- **Fisión multimodal mediante el uso de estándares**. Sería deseable el desarrollo de un modulo específico de "fisión" multimodal. Al igual que el módulo de "fusión" se encarga de unir las entradas sensoriales en *actos comunicativos*, que abstraen al gestor del diálogo del canal o los canales (modos) por los que se ha obtenido dicha información; el componente de "fisión" se encargaría de abstraer al gestor del diálogo de los canales/modos usados para expresar un cierto mensaje (*actos comunicativos de expresión* o *acciones comunicativas*).

  Hasta ahora, en la especificación del diálogo (en forma de fichero VoiceXML), es necesario detallar cada uno de los modos con los que el sistema desea expresarse. Por ejemplo para hacer un saludo, es necesario indicar el texto a decir, los gestos a realizar, los sonidos a generar, etc.

  Con un módulo de fisión multimodal se logra un doble objetivo. Por un lado el de simplificar la especificación de los diálogos interpretados por el gestor del diálogo. Por otro lado, facilitar la reutilización de estas acciones comunicativas en diferentes diálogos.

- **Mejora del sistema de identificación de usuarios**. En el trabajo aquí presentado la identificación de usuarios se realiza meramente usando el tono de voz (canal sonoro). Se está trabajando, dentro de nuestro grupo, en el desarrollo de un sistema bimodal (visión y voz). El tono de voz, la identificación de la cara, la detección del género, así como el color de la vestimenta pueden servir para mejorar la precisión de la identificación.

- **Mejora del sistema de detección de emociones**. Sería interesante añadir al sistema de detección de emociones, información relevante al contexto comunicativo y a la historia previa del diálogo. Si el diálogo presenta una alta tasa de errores de reconocimiento de voz, probablemente el usuario este más triste o aburrido, que feliz. Al igual que, si en la conversación aparecen palabras como feliz, divertido, etc, es muy probable que el usuario se encuentre en un estado de felicidad, mientras que si aparecen palabras como cansado o desanimado, muy probablemente el usuario se encuentre aburrido. Siguiendo esta aproximación, que tiene en cuenta la historia previa del diálogo con el usuario, tendría sentido estudiar las emociones generadas en el usuario por el propio proceso de diálogo: aburrimiento, diversión, duda...

- **Interpretación pragmática del diálogo**. El sistema presentado trabaja a un nivel semántico, sin tener en cuenta el aspecto *pragmático* del discurso. En este nivel se debería ser capaz de discernir la intención del mensaje en su contexto. Una misma frase dicha en contextos diferentes, puede tomar significados completamente diferentes, especialmente si la frase es irónica.

Dentro de este nivel pragmático, se hace necesario una "mayor comprensión" de los mensajes intercambiados en el diálogo. En ese sentido, es necesario establecer relaciones entre la información intercambiada en la interacción con información proveniente del mundo real. Se está trabajando en esta vía de investigación, dando lugar a lo que hemos llamado sistema aumentado de diálogo robótico (ARDS).

- **Superar completamente las dificultades en el reconocimiento de voz**. En el capítulo donde se describe el reconocimiento automático del habla, se presenta el "Cocktail Party Problem", y los desafíos en áreas como la "cancelación activa del ruido", "cancelación del eco", "detección de actividad de voz". Para lograr un sistema en el que la interacción sea completamente satisfactoria, es necesario hacer un esfuerzo importante en el desarrollo de algoritmos y hardware que permitan una interacción natural (full-duplex).

- **Inclusión de una herramienta de "ChatBot"**. Permite interacción natural no sujeta a ningún contexto específico. Lo más parecido que se ha desarrollado para este trabajo es el diálogo de pregunta abierta, en el que el usuario puede formular una pregunta y el robot le responde consultando en bases de datos semánticas. Esta nueva herramienta permitiría conversar sobre multitud de temas.

11.3. Publicaciones surgidas de este trabajo

El desarrollo de este trabajo científico ha dado lugar a la generación de literatura que se enumeran continuación.

11.3.1. En revistas científicas

1. Alonso-Martin, F.; M.Malfaz; J.F.Gorostiza; M.A.Salichs. A Multimodal Emotion Detection System during Human–Robot Interaction. Sensors (Online). Vol. 13. No. 11. pp.15549-15581. 2013.

2. Alonso-Martin, F.; J.F.Gorostiza; M.Malfaz; M.A.Salichs. Multimodal Fusion as Communicative Acts during Human–Robot Interaction. Cybernetics and Systems: An International Journal (Online). Vol. 44. No. 8. pp.681-703. 2013.

3. Alonso-Martin, F.; J.F.Gorostiza; M.Malfaz; M.A.Salichs. User Localization During Human-Robot Interaction. Sensors (Online). Vol. 12. No. 7. pp.9913-9935. 2012.

4. Alonso-Martin, F.; M.A.Salichs. Integration of a voice recognition system in a social robot. Cybernetics and Systems: An International Journal (Online). Vol. 42. No. 4. pp.215-245. 2011.

5. Alonso-Martin, F.; A.Ramey; F.Alonso; A.Castro-González; M.A.Salichs. Maggie: A Social Robot as a Gaming Platform . International Journal of Social Robotics (Online). Vol. 3. No. 4. pp.371-381. 2011.

6. Alonso-Martin, F.; A.Castro-González; J.F.Gorostiza; M.A.Salichs. Augmented Robotics Dialog System for Social Robots. Sensors (Online) (Pendiente de publicar 2014)

7. J.F.Gorostiza; Alonso-Martin, F.; M.A.Salichs. Quasons: Versatile Electrosonic Elements for Expressing Meaningfull and Affective Sounds for Natural Interaction with Social Robots. (Pendiente de publicar 2014)

11.3.2. En congresos de robótica

1. Alonso-Martin, F.; Ramey, A.; Salichs, M. Á. (2014). Speaker identification using three signal voice domains during human-robot interaction. In Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14 (pp. 114–115). New York, New York, USA: ACM Press.

2. Alonso-Martin, F.; A.Castro-González; J.F.Gorostiza; M.A.Salichs. Multidomain Voice Activity Detection during Human-Robot Interaction.. International Conference on Social Robotics (ICSR 2013). . Bristol. UK. Oct, 2013.

3. Alonso-Martin, F.; J.F.Gorostiza; M.A.Salichs. Descripción general del sistema de interacción humano-robot Robotics Dialog System (RDS). RoboCity2030 12th Workshop: Robótica Cognitiva. Madrid. Spain. Jul, 2013. . ISBN: 9-788469-58175. UNED. pp.0-0. 2013.

4. Alonso-Martin, F.; J.F.Gorostiza; M.A.Salichs. Preliminary Experiments on HRI for improvement the Robotic Dialog System (RDS). Robocity2030 11th Workshop: Robots Sociales. Leganés. España. Mar, 2013.

5. Alonso-Martin, F.; J.F.Gorostiza; M.A.Salichs. Musical Expression in a Social Robot. Proceedings of the 2012 International IEEE Intelligent Vehicles Symposium. Workshops V Perception in Robotics. Alcalá de Hena. Spain. Jun, 2012.

6. Alonso-Martin, F.; A.Ramey; M.A.Salichs. Maggie : el robot traductor. . 9º WorkShop Robocity2030. Madrid. Spain. May, 2011.

7. Alonso-Martin, F.; V.Gonzalez; A.Castro-González; A.Ramey; M.A.Salichs. Using a Social Robot as a Gaming Platform. International Conference on Social Robotics 2010. Singapore. Nov, 2010.

11.4. Comentarios finales

A medida que se ha ido profundizando en la investigación y el campo del conocimiento se ha ido ampliando, nuevas puertas se han ido abriendo. Esta tesis no se considera un punto y final en el que el proceso de investigación concluye. Se espera que constituya un punto y seguido, abriendo una nueva etapa en la que nuevos avances supongan una mejora sustancial en el trabajo aquí expuesto.

Ha sido difícil tomar la decisión de en qué momento fijar este punto y seguido al trabajo realizado. Muchos de los avances que están a punto de ver la luz han sido descritos en el apartado anterior.

CAPÍTULO 12

Conclusions

12.1. Conclusions

A doctoral thesis is comprised of a research process in a specific problem of a particular scientific area, and a report at the end of it. This report is created to describe the state-of-the-art of the problem, what has been the process to study it, the results obtained and what they mean. Finally future research lines are mentioned as continuation of this work and to enhance it. Thus, this work can be considered as a doctoral thesis. It is located on the human-robot interaction research field. Its main challenge is to improve the current human-robot interaction systems to make them more similar to the ones existing among human beings. Thus, to reach to this point, a general interaction system has been proposed initially and then, it has been divided into several smaller problems. To solve each one, a similar process has been followed: analyse the state-of-the-art, describe the research process and finally show the new results.

- General interaction system whose name is RDS. It works with several interaction input and output modes by mixing the multi-modal inputs and splitting the multi-modal outputs. Among these input modes, the ones related to the sound analysis are especially important (multi-sound system); they enable to use the audio input for the following tasks: voice recognition, user identification, spatial location of the user in relation with the robot and user emotion detection. Among the output modes, the important ones are: emotional voice synthesis, non-textual real-time sounds and musical skills. This sound mode is complemented with the rest of the modes: visual, gesture, radio-frequency and tactile modes.

- The multi-modal input is managed according to the communicative act theory between human beings. Thus, the “Interaction management system" is not dependent on the modes or channels where the message has been captured from. Each detected communicative act is comprised of several multi-modal attribute-value pairs. A new subcomponent has been developed to aggregate this timed semantic values in a coherent form – information packages belonging to a interaction turn.

- The multimodal input is managed inspired on the theory of communicative acts of the interaction between humans. Thus, the dialog manager receives the communicative messages regardless the modes used to transmit them. Each communicative act perceived, is an information package with attribute-value pairs. The module that package this information is called multimodal fusion.

- he interactivity system is completely customisable by each user; this is achieved by generating, storing and using user profiles by the natural interaction system.

These profiles have information regarding the user preferences as follows:

1. Language adaptation (multi-lingual): the system is able to speak in the language the user is using in the interactions. This adaptation is valid for all available languages, but works really good for English and Spanish. It's only necessary that the user greets the robot to find out the user and the language that he/she is using in the communication.
2. Interaction distance customisation (proxemics): the system can find the interaction distance for each user or group of users.
3. Familiarity adaptation: provided that there is a previous user experience with the system the robot is able to customise each interaction. In other words, a novel user can be "taught" by the robot itself, while an experienced user only needs short precise conversations about some questions or suggestions.
4. Emotional customisation: the system can customise the interaction to match the user emotions captured by the system, as well as the emotional state of the robot itself 1.

After having reviewed all the general contributions, the main detailed contributions found for each component will be analysed and described in depth. Furthermore, each detailed contribution is included inside different, independent research fields.

- A new voice recognition system has been developed to work with two different and complementary modes at the same time: the first based on grammars and semantics, and the second one is a free natural language mode based on language-dependent statistical models (see chapter 6 and Appendix C). On the other hand, different software and hardware configurations, which were designed to achieve this task in a satisfactory way in human-robot interactions, have been analysed as well as the problems that may arise in the process.

- A multi-modal robot-centric user location system has been developed, it has been integrated inside the proposed interaction system. It also includes a set of rules obtained from a systematic proxemics study between our robot Maggie and some users which enables the robot to adapt the interaction with a concrete user or a group of users.

- A conversation management system based on filling information gaps has been added, this system is based on a previous doctoral thesis and the VoiceXML standard. It has also been extended with additional multi-modal features that are not included in this standard, but they were available in the robot, such as

multi-modal input and output, adapting to several languages, semantic queries to online web services, playing music by using online music services, etc.

- A verbal and non-verbal voice recognition system has also been developed. It has been designed in a multi-layer architecture. The layers included in the system are: abstraction layer, voice and sound recognition with several voice recognition systems working at the same time (currently six systems). Some other features has been included, such as voice-recordings queue management. The non-verbal voice recognition system produces sounds which emulates emotions from the robot itself as well as copying the user voice but with the robot language.

- And a fully-integrated multi-modal emotion detection system has been developed within the interaction system. It is used as an input – it is able to detect emotions, by analysing the user voice, and his/her facial expressions, by using artificial vision techniques. This emotional awareness is essential to customise the robot interactions and avoid general failure with the user.

This task not only has been a research process, but also, a great effort has been done in integrating tasks in order to have a long-term reliable system.

12.2. Current and future work

In this chapter some improvements and ongoing tasks will be described:

- Multi-part conversations: the system is currently designed to develop full-duplex interactions with one user at a time. It's not possible to have a coherent simultaneous interaction between several users and the robot. To achieve this goal, there are some challenges to face, such as:

    1. Split the different channels for each mode. Each mode shall have as many channels as users are present in the communication. For example, the voice recognition system shall receive every user voice through a different channel.
    2. Advanced conversation-phase management. The system shall be able to decide the target user for each interaction.
    3. Taking into account several user profiles. The system shall be able to load all the user profiles for each user included in the conversation. It shall be able to alternate among them as required.

- Multi-modal fision using standards. It would be necessary to integrate a specific module to manage it. As it has been done in the fusion module (where all the sensorial inputs are gathered and included in communicative acts to abstract the dialog manager from the input channel), the splitting should be able to abstract the dialog manager from the output channels. Currently, in the dialog specification (VoiceXML file), it's necessary to describe each mode the system needs to interact. For example, to greet the user, it's necessary to set the output text, the gestures, the sounds to generate, etc. If this multi-modal splitting were available, the VoiceXML specification would be simplified and it would also be an improvement if it would have the possibility to reuse some communication act between conversations.

- Improvements in the user identification system. In this work, the user identification is only based on the voice tone (detected in the sound input). One of our current research lines is to develop a bimodal system which, not only takes into account the voice, but also the vision. The voice tone, the face identification, the gender detection as well as the user clothing are also valid to enhance the identification system precision.

- Improvements in the emotion detection system. It would be useful to add new contextual information in the conversation as well as some previous steps already done in it. If the conversation had a high level of errors in the voice recognition, it would turn out in a bored or sad rather than happy user. Other possibility is that if in the conversation speeches appeared words such as happy, funny, etc., it would be very likely that the user may be happy, while, on the other hand, if there would be other words such as tired or unhappy, the user may be bored. With these premises and the previous steps, a successful conversation can be achieved.

- Pragmatic interpretation of the conversation: the system presented in this thesis is a semantic-based one. Thus, it is ignoring pragmatic point of view of the conversation. Then, a major upgrade would be taking into account this pragmatic context. This would allow to infer the intention of a message; the same sentence can have complete different meaning, for example, as a result of an irony. At this pragmatic level, it is necessary a more precise interpretation of the silences in the conversation. Silence can be shown as a communicative act and it should be managed with a suitable semantic meaning, for example, when a user has not understood, or he could not hear it, or he may have some doubts about the next action to do, etc. On the other hand, other silence moments have the goal of breathing and getting enough time to think ahead of the next steps to take in the conversation. At this point, , a much more complex model than the one

presented in this dissertation is needed.

- To overcome the difficulties found in the voice recognition system. The "Cocktail Party Problem" and other challenges in other fields such as the "active noise cancellation", "echo cancellation" and "voice activity detection" have been described in the chapter focused on the automated voice recognition system. Solving these problems would lead to a fully satisfactory interaction. It would also include a remarkable effort in new algorithms and hardware to allow a natural (full-duplex) interaction.
- Adding a "ChatBot"-like tool. It would allow a context-less natural interaction. The most similar work included in this work is the open-ended dialog. In this dialog, the user queries the robot and it answers based on results retrieved from opened semantic databases. This new tool would allow conversations dealing with many different matters.

12.3. Last comments

As the research had been progressing, the field has been widened and new possibilities have appeared. This thesis cannot be seen as an ending point for this research, but only a milestone to open a new stage in which new features effectively add a substantial improvement in the work done in this thesis. It has been really difficult to make a decision where to set a full stop in this work. The vast majority of the improvements described in the last point will be released briefly.

APÉNDICE A

Primeros resultados experimentales con el sistema RDS

A.1. Introducción

En este apéndice se presentan los experimentos realizados con usuarios reales, para la evaluación del sistema de diálogo RDS. Primeramente se hace una descripción de las cuatro series de experimentos realizados: uno con adolescentes de manera individual (de unos 13 años), y los otros tres con pequeños grupos de chicos (de alrededor de 8 años). Se han usado dos métricas: análisis de vídeo y cuestionarios. En el análisis de los vídeos, se han elegido un conjunto de parámetros propios de la interacción, para su grado de efectividad y facilidad de uso del sistema. Posteriormente se muestran los resultados obtenidos de estos análisis, en un conjunto de parámetros relacionados con la percepción subjetiva de la facilidad de uso, diversión, efectividad, etc.

La realización de estos experimentos con usuarios reales, suscitó la introducción de numerosas mejoras en el sistema de interacción que se enumeran en la parte final de este capítulo.

A.2. Descripción de los experimentos

En todos los casos, se quería medir la facilidad de uso, lo intuitivo y efectivo que es el uso del sistema RDS con los diálogos implementados y que permiten interaccionar con el robot: en que aspectos comunicativos el sistema se comporta como aburrido o divertido, la versatilidad del dialogo en diferentes contextos comunicativos y su capacidad de ayuda/recuperación frente a errores.

El método de evaluación está basado en el análisis de algunas escenas con interacción entre el robot y el o los humanos, que fueron grabados en vídeo y posteriormente sometidos a rellenar unos cuestionarios. En estos cuestionarios, los usuarios respondieron a algunas preguntas, con valores numéricos comprendidos entre 0 y 10 (para adolescentes) y entre 0 y 4 (para niños). A través de esas preguntas, se quería determinar como el usuario percibe la interacción con el robot, que fue representada con las siguientes características:

- **Facilidad de Uso**. Como de fácil es de utilizar e interaccionar con el robot sin conocimiento previo sobre el robot y el sistema. Esto depende de si el sistema es lo suficientemente intuitivo.
- **Entretenido/Divertido**. Un valor sobre el grado de diversión que el usuario siente cuando interactúa y juega con el robot.
- **Inteligencia**. Este valor da una idea sobre la percepción que tiene el usuario del grado de inteligencia del robot.
- **ASR**. Grado de entendimiento que tiene el robot sobre lo que el usuario le dice.

- **TTS**. Grado de comprensibilidad y expresividad de la voz del robot.
- **Comportamiento**. Si el comportamiento del robot ha sido coherente durante la interacción. Esto quiere decir, que los actos del robot han sido coherentes con lo que el usuario le ha demandado.
- **Naturalidad**. Este aspecto trata de medir, como de natural se comporta el robot, o como de parecido es al comportamiento humano. Este parámetro está bastante relacionado con la facilidad de uso.
- **Aceptación**. Con este parámetro se quiere medir el grado de aceptación del robot y de la interacción, la pregunta concreta es si el usuario prefiere al robot Maggie o sus dispositivos electrónicos favoritos (vídeo consolas, ordenador, etc).
- **Sensación producida**. Sensación general después de la interacción con el robot, da satisfacción o de frustración.

Analizando los vídeos, se han recopilado datos como el tiempo máximo de interacción de cada usuario con el robot antes de cansarse o aburrirse, la efectividad de la comunicación (fallos en el reconocimiento de voz o en la expresión) y el funcionamiento general de todo el sistema (numero de habilidades activadas, tipo de habilidades activadas, número de subdiálogos de aclaración, etc).

Todos los experimentos han sido hechos on-line, ejecutando la arquitectura de control al completo, que integra todas las capacidades del robot, por lo que el robot ha actuado de forma totalmente autónoma. Previamente a estos experimentos, se había usado métodos de "Mago de Oz", pero solo para hacer un ajuste fino de las gramáticas usadas, los gestos, etc.

En esta sección, se describen los casos de estudio realizados:

A.2.1. Caso I: interacción individual con adolescentes inexpertos

En este primer caso, se ha experimentado con un total de 7 chicos (2 chicos y 5 chicas) de entre 13 y 16 años. Cada chico entró individualmente en el laboratorio, donde Maggie les estaba esperando (ver Fig. A.1). Esté fue su primer encuentro entre ambas partes. Se colocó a cada usuario un micrófono inalámbrico [1] y se le dijo que actuara con naturalidad. El sistema ha sido diseñado para que en el caso que el robot no conozca al usuario sea auto-explicativo. En el caso de fallos de reconocimiento de voz, el robot sólo pregunta por la información precisa que necesita.

[1] Recientemente ya no es necesario el uso de micrófono inalámbrico por parte del usuario, ya que el propio robot tiene incorporados sus mecanismos para captar el sonido

Figura A.1: Interacción individual con adolescentes sin experiencia previa con el sistema

Como se había notado en otros experimentos, para los niños Maggie es muy atractiva, por lo que algunas veces es muy difícil medir de que depende que pueda llegar a aburrirse. En otras palabras, como medir el grado de diversión. Para dar una idea de la influencia de la perdida de interés en la comunicación (sensación de aburrimiento), se ha medido el tiempo de interacción total del usuario con el robot. Este parámetro nos puede dar una idea del grado de diversión en la interacción. Para medir el grado de diversión también se ha tenido en cuenta el resultado de los cuestionarios.

En algunos casos el usuario se encuentra perdido y mira al desarrollador (nosotros) pidiéndonos ayuda. Entonces se interviene dándole algún ejemplo de uso. Después se vuelve a dejar al usuario que continue interactuando con el robot hasta que el quiera.

Se ha tomado un vídeo de cada usuario interactuando. De esta manera se ha podido medir algunos parámetros dinámicos como el tiempo total de interacción, se había momentos de largo silencio (más de 2 segundos), número de turnos intercambiados, fallos de ASR, momentos en los que el usuario recibía ayuda, perdidas de coherencia, etc.

A.2.2. Caso II: interacción en pequeños grupos de niños inexpertos

En este segundo caso, los experimentos se han llevado a cabo con un pequeño grupo de cinco niños entre 8 y 9 años (ver Fig. A.2). También se ha querido comprobar como de fácil es establecer la comunicación si el usuario no sabe nada sobre el robot y es el robot el que explica el mismo que es capaz de hacer y como es necesario interactuar con él.

Nuevamente se ha dado un micrófono inalámbrico al grupo de niños, para que el robot reciba la señal de voz del usuario que quiera hablar. El resto de niños podían acercarse al robot y tocarle, coger el micrófono y tomar la iniciativa o sugerirle posibilidades al niño poseedor del micrófono.

El robot incorporó todas sus funcionalidades: comunicación verbal y no verbal (asr, etts, gestos, sonidos, localización de la fuente sonora, etc), y habilidades especiales para el control de la televisión por diálogo, juegos con peluches, lector de noticias,

habilidad de karaoke, etc.

Figura A.2: Grupo de niños interactuando con el robot sin ayuda

Y finalamente, todos los usuarios respondieron a los mismos cuestionarios. También se analizaron los vídeos para tomar conclusiones sobre la interacción. En este caso se hizo un análisis cualitativo, puesto que la interacción en grupos es más difícil de analizar cuantitativamente.

A.2.3. Caso III: Interacción en pequeños grupos con niños supervisados

En este caso, se ha tomado otro pequeño gropo de niños, como en el caso anterior, pero previamente le se ha dado unas pistas y una demostración de como interactuar con Maggie (ver Fig. A.3). Por ejemplo, enseñándoles como hablar, que comandos de voz entiende mejor, el repertorio total de habilidades del robot, etc. Esta vez se ha focalizado más en las características que tienen que ver con la interacción no verbal.

Por lo tanto, una vez que les se ha hecho un breve tutorial de como interactuar con el robot, se le ha dado el micrófono a uno de los niños y se le ha dejado interactuando con el robot. Nuevamente se ha grabado todo en vídeo y entregado al final unos cuestionarios.

Figura A.3: Investigador presentando a Maggie

A.2.4. Caso IV: interacción en grandes grupos con niños supervisados

En este caso, más de 25 niños interactuaron simultáneamente con el robot. Este experimento se focalizó en estudiar la interacción verbal y no verbal que ofrece el sistema de diálogo RDS, y su relación con la facilidad de uso y el entretenimiento que posibilita el sistema en si mismo (ver Fig. A.4).

Cualquiera de los niños podía coger el micrófono y hablar al robot y cualquiera podía jugar con los juegos que el usuario activaba para jugar con el robot. Además se han tomado vídeos de la interacción caótica producida por 25 niños de 8 años, dentro del laboratorio intentando actuar/jugar/pegar con el robot ...

Figura A.4: Interacción en grandes grupos con niños supervisados

A.3. Análisis de los vídeos

En esta sección se presentan los resultados obtenidos de análisis de los vídeos de diferentes escenarios de interacción descritos en la sección superior.

A.3.1. Caso I: interacción individual con jóvenes inexpertos

En la tabla A.3.1 se muestran algunos parámetros descritos, como la coherencia y efectividad en la interacción para el Caso I.

Se puede resumir los resultados como sigue. El tiempo total de interacción en los diferentes caso varia entre 2 y 12 minutos usando entre 1 y 4 habilidades diferentes. Esto indica que se usa cada habilidad 3 minutos, que está dentro de lo que se podría esperar y está también dentro de lo que se puede considerar un uso normal de cada habilidad. En otras palabras, el uso de diferentes habilidades no necesita mucho tiempo de aprendizaje, puesto que los mecanismos de interacción son los mismos.

Respecto al análisis de los turnos, nos gustaría subrayar lo siguiente. Se detecta unos 4 turnos intercambiados por minutos. Este valor es muy bajo en comparación con

	Media	Desviación	Rango
Número de turnos	25	±13	[12, 48]
Número de frases	25	±11	[12, 49]
Fallos de reconocimiento	13	±5	[8, 30]
Frases del robot	35	±22	[6, 66]
Tiempo total (mins)	7	±3	[4, 12]
Momentos de silencio	5	±3	[2, 13]
Tiempo de silencio (secs)	27	±14	[9, 53]
Num. Ayudas de los expertos	10	±6	[2, 26]
Habilidades Usadas	2	±1	[1, 4]

Cuadro A.1: Resultados tras en análisis de los vídeos del Caso I

una interacción natural y recalca la idea que la interacción fue lenta. En parte es causado por el hecho de que el robot da demasiadas instrucciones y habla mucho el mismo, dado que no conoce al usuario y se encuentra en modo "auto-explicativo/tutorial". Este hecho es también corroborado con el numero de palabras por turno que es mayor por parte del robot que por parte del usuario, ya que el usuario es mucho más conciso. De hecho, en todos los casos los usuarios expresan sólo una frase en cada turno. Esto muestra que los usuarios ven al robot como una máquina a la que deben hablar con mucho cuidado (sólo una frase cada vez), porque no es capaz de entender lo que el humano hace. Los usuarios notan que algunas veces el robot tiene dificultades para entender alguna cosa que es dicha, y rápidamente lo tienen en cuenta para adaptar su forma de hablar a una frase por cada turno.

En la tabla A.3.1, se da mas datos específicos sobre el reconocimiento de voz, el rol del silencio, y los momentos en los que el experto toma la iniciativa para ofrecer algo de ayuda al usuario.

	Media	Desviación	Rango
Precisión del ASR	46 %	±12 %	[17, 70]
Silencio vs Tiempo total	10 %	±7 %	[4, 25]
Momentos de silencio vs Turnos totales	22 %	±9 %	[8, 33]
Momentos de ayuda vs. Turnos totales	25 %	±10 %	[5, 40]

Cuadro A.2: Valores relativos interesantes del análisis de los vídeos para el Caso I)

Respecto al reconocimiento de voz, el valor medio fue calculado sobre las frases reconocidas respecto del total de frases pronunciadas por el usuario. Se ha obtenido un valor muy bajo en comparación con experimentos previos (como por ejemplo en [Alonso & Salichs, 2011]). La causa corresponde a dos orígenes. El primero de todos,

es que los usuarios no tenían ningún conocimiento sobre las gramáticas que usa el reconocedor de voz, que es lo mismo que decir que ellos no conocían nada sobre el lenguaje formal que el robot es capaz de reconocer. Por lo que el reconocimiento falla cuando los usuarios dicen frases y palabras fuera del vocabulario/gramáticas que el robot es capaz de interpretar en eses contexto. Sin embargo, esto ha servido, para aumentar notablemente el conjunto de reglas recogidas en cada gramática, con todas las expresiones recogidas y analizadas en los vídeos, con la consiguiente mejorar en la tasa de reconocimiento para futuros experimentos. La otra causa principal de fallos de reconocimiento, es que los niños/jóvenes suelen expresarse con menos firmeza (estabilidad en la voz), que un adulto, especialmente si es la primera vez que interactúan con un robot.

A pesar de que la mitad de las freses no fueron reconocidas correctamente, los escenarios de interacción fueron lo suficientemente fluidos, de hecho el número de momentos de silencio, donde ni el robot ni el usuario sabían que decir (perdida de coherencia), es relativamente baja, sólo el 22 % del total del turnos y menos del 10 % del tiempo total de interacción.

Algunas veces, el investigador/desarrollador tomó la iniciativa, sólo para mostrar al usuario que debía decir exactamente al robot, para que este último le entendiese. Esos breves momentos representaron aproximadamente el 25 % del total de turnos, un valor muy similar al de los momentos de silencio. Por lo que la cantidad de veces que el desarrollador ayudó al usuario corresponde con las veces que se perdió completamente la coherencia.

A.3.2. Caso II: interacción en pequeños grupos con niños inexpertos

En este caso, no se ha hecho una medida cuantitativa de cada parámetro relevante, sin embargo se ha hecho un análisis cualitativo. Esto es porque los experimentos se hicieron de un modo diferente al caso anterior. Por ejemplo, el reconocedor de voz no funciona tan bien para niños de esas edades, y la interacción no verbal toma mayor protagonismo.

En este escenario, una chica y el robot intentan mantener una conversación. Sin embargo, el reconocedor de voz falla creando gran hilaridad en los chicos. Ellos se aproximan al robot e intentan otros modos para interactuar con él: los modos de visión y tacto. La niña también expresa su estado de ánimo: es la primera vez que interactuaba con el robot y estaba nerviosa, etc. Después de un minuto, la chica y el robot lograr entablar una conversación coherente: el robot le pregunta por su nombre, la saluda y la pregunta por más información personal como su idioma preferido, edad ... información que usa para crear un perfil de usuario de esa niña.

Los otros niños se colocan a una distancia de 3 metros del robot y parecen pasárselo

muy bien, ya que están continuamente riéndose. Algunas veces se aproximan al robot y sugieren cosas que la niña que tiene el micrófono debe decir, siempre entre muchas risas y carcajadas.

En otro escenario, otro chico coge el micrófono e intenta también interactuar con Maggie, pero nuevamente el reconocimiento de voz defectuoso pone trabas a la interacción natural. A pesar de ello el robot le sugiere al niño que hable mas bajo y mas lentamente, sin embargo el niño grita al robot acrecentando las risas del resto de sus compañeros, todo ello en un ambiente muy amigable y divertido. En general, los niños se lo pasaron muy bien y no pidieron, en ningún momento, a los investigadores parar el experimento.

La principal causa de fallos en la interacción fue la carencia de adaptación del modo de audio, Por ejemplo, el sistema no adaptó su sensibilidad de reconocimiento al nivel del volumen de la voz y ruido del entorno. Debido a estos experimentos se ha trabajado en solventar estos problemas por varias vías. Una de ellas es, para interacciones con niños pequeños, basar la interacción en etiquetas de radio frecuencia, con ilustraciones/pictogramas de lo que quiere el niño que haga el robot (ver Fig. A.5). Por otro lado, se ha mejorado la interacción con voz, mediante un control activo de volumen (por hardware), que permite normalizar el volumen de entrada de audio, sea cual sea su intensidad; tanto si el usuario habla alto como bajo, al reconocedor de voz le llega una señal de audio de un volumen adecuado. Otras técnicas, como la cancelación activa del eco y la cancelación de ruido estacionario contribuyen a mejorar la interacción por voz.

Figura A.5: Tarjetas con pictogramas y etiquetas de radio frecuencia

A.3.3. Caso III: interacción supervisada en pequeños grupos de niños

Después de hacer un mini tutorial sobre como usar el robot y darle el micrófono a uno de los 5 chicos, ellos comenzaron a hablar con el robot. En este caso, el reconocedor funcionó muy bien, y la coherencia se mantuvo estable. Primeramente el robot les preguntó por su edad, nombres, etc. para a continuación darles algunas instrucciones de como interactuar con el mismo. Seguidamente los niños, consiguieron interactuar con el robot con éxito activando ciertas habilidades, juegos y acciones que el robot es capaz de desempeñar. Habilidades como cantar, bailar, controlar la televisión, jugar a juegos de adivinar personajes, perseguir a una persona, etc. Los chicos aprendieron también muy rápido como detener y activar cada una de las habilidades.

Un juego que especialmente le gusto a los chicos fue el de jugar con los peluches (ver la Fig. A.6). Un juego especialmente diseñado para chicos de esa edad y que consiste en reconocer formas, figuras, animales, idiomas, todo ello fomentando el trabajo en grupo. La interacción durante el juego se base en voz y etiquetas de radio frecuencia. El juego consiste en que el robot les hace una pregunta relativa a algún peluche de los disponibles, los niños deben buscar el animal/peluche que creen que mejor responde a dicha pregunta y se lo acercan a su nariz. El robot identifica dicho peluche mediante el sensor de radio frecuencia y les dice si la respuesta es correcta o incorrecta. Esta operación se repite hasta que se completan un determinado número de preguntas/respuestas (para una mayor explicación consultar [Gonzalez-Pacheco et al., 2011]). Jugando a este juego se han encontrado algunas características especiales dignas de comentar. Lo primero de todo es que todos los chicos se mostraban muy activos e involucrando durante toda la interacción, lo segundo es que empiezan a autoorganizarse, de tal manera que cada uno de ellos desempeña un papel especial en el juego. Por ejemplo, algunos chicos repetían al resto lo que el robot había preguntado, otros buscaban de entra la pila de peluches el que creían que era el adecuado, mientras que otro niño (el mas alto), se encargaba de pasárselo por la nariz del robot. Esta emergente y espontánea actitud colaborativa entre los niños y el robot nos ha llamado tremendamente la atención.

Otra actividad que surgió y vale la pena destacar, es la siguiente. Cuando el robot activa su habilidad de cantar y bailar, los chicos comenzaron espontáneamente a imitar los movimientos del robot. Se colocaron juntos en una fila e intentaron aprender la coreografía que el robot estaba llevando a cabo. Este hecho demuestra que los robots sociales podrían ser muy útiles para una clase de gimnasia o baile u otra actividad donde los niños tienen que repetir e imitar los movimientos de un maestro/líder.

Figura A.6: Niños jugando al juego de los peluches

A.3.4. Caso IV: interacción supervisada en grandes grupos de niños

En este experimento, se nota que el grado de diversión se incrementa proporcionalmente al número de chicos (unos 255). En el grupo, había chicos que ya conocían como se usaba el robot puesto que ellos habían estado presente en algunos experimentos. Por lo tanto, los roles entre ellos fueron fácil y rápidamente repartidos: uno de ellos con experiencia previa cogía el micrófono y guiaba al resto, y los otros jugaban a los diferentes juegos que el "maestro" activaba. Observando la manera de interactuar con el robot, rápidamente el resto de chicos también se unieron a la interacción.

También se ha observado que algunos chicos directamente no prestan atención al robot. Por ejemplo, algunos comenzaron a hablar con los investigadores del laboratorio, preguntándoles si podían usar un patinete presente en la sala, o sólo pidiendo si podían usar un ordenador de los también presentes. Esto podría ser debido al hecho de que los chicos habían estado visitando la universidad antes de iniciar el experimento y podrían estar ya más cansados o saturados que otros chicos.

A.4. Análisis de los cuestionarios

A.4.1. Caso I: interacción individual con adolescentes inexpertos

Lo primero de todo, los cuestionarios intentan dar una idea de como de habituados están los chicos a interactuar con videojuegos y ordenadores, para darnos una idea del grado de familiaridad y atracción por los dispositivos electrónicos. Todos ellos usaban habitualmente esos dispositivos entre 5 y 10 horas por semana, lo que puede ser considerado como un rango "normal" en nuestra sociedad [2].

[2] En un estudio hecho por la sociedad española FAD (Fundación de Ayuda a la Drogradicción) en 2002, mas del 40 % de los adolescentes (en edades comprendidas entre los 14 y 18 años) juegan

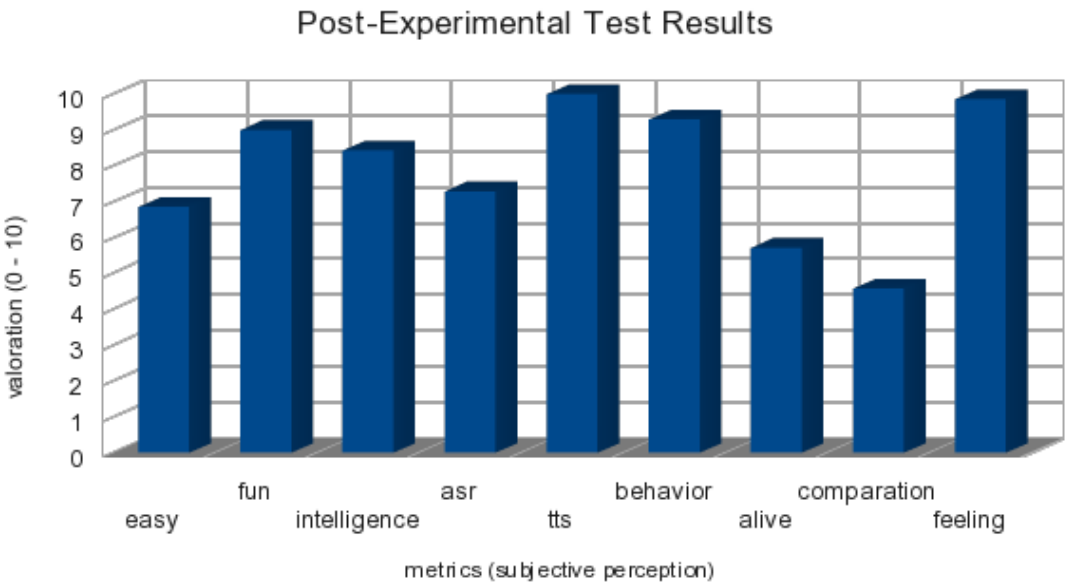


Figura A.7: Caso I. Después de que algunos adolescentes inexpertos actuaran con Maggie individualmente. Cada columna representa valores medios en el rango de 0-10, como resultado de las cuestiones sobre algunos aspectos subjetivos de percepción de cada experimento.

En la Fig. A.7 representa los resultados de los tests. En cada columna se representa los valores medios dados por el usuario. Todos los usuarios califican la interacción como una experiencia satisfactoria (valor de la `sensación producida`). Pero el sistema todavía carece de facilidad de uso. De hecho, durante los experimentos algunas veces el desarrollador tubo que dar algunas pistas somo como usar el robot y ampliar las explicaciones que el robot daba naturalmente, guiando al usuario en el modo correcto de decir alguna cosa (con las palabras específicas que el robot era capaz de entender).

No obstante, los usuarios percibieron al robot como una máquina inteligente. Su comportamiento era coherente con la interacción (cuando los malentendidos sucedían el robot reaccionaba adecuadamente recobrando la coherencia en la interacción). Hay que tener en cuenta que aunque los mal entendidos podían ocasionarse fundamentalmente por fallos en el reconocimiento de voz, muy rara vez ocurrían casos de falsos positivos, es decir, que el robot entendiese algo diferente a lo dicho. Si el robot entendía algo, casi siempre coincidía con lo que el usuario había dicho. Por lo tanto, a pesar de los errores en dicho reconocimiento de voz (algo muy habitual en interacción

juegos electrónicos entre 1 y 2 horas al día en días laborables, y mas del 26 % juegan mas de 3 horas al día los fines de semana.

natural entre humanos) el diálogo era capaz de reaccionar adecuadamente. Lo que el robot expresaba por voz y por gestos también fue fácilmente entendido. Finalmente los usuarios percibieron la interacción como divertida y satisfactoria. Sin embargo, cuando se les pregunta, en general, si preferían el robot sobre su ordenador personal o videoconsola, la respuesta general fue que preferían sus dispositivos electrónicos frente a Maggie.

A.4.2. Caso II: interacción en pequeños grupos con niños inexpertos

En los casos donde los usuarios tenían al rededor de 8 años de edad, usan ordenadores y juegos electrónicos durante menos de 5 horas a la semana. Este resultado puede ser considerado un poco por deajo del rango normal de los niños españoles descrito superiormente.

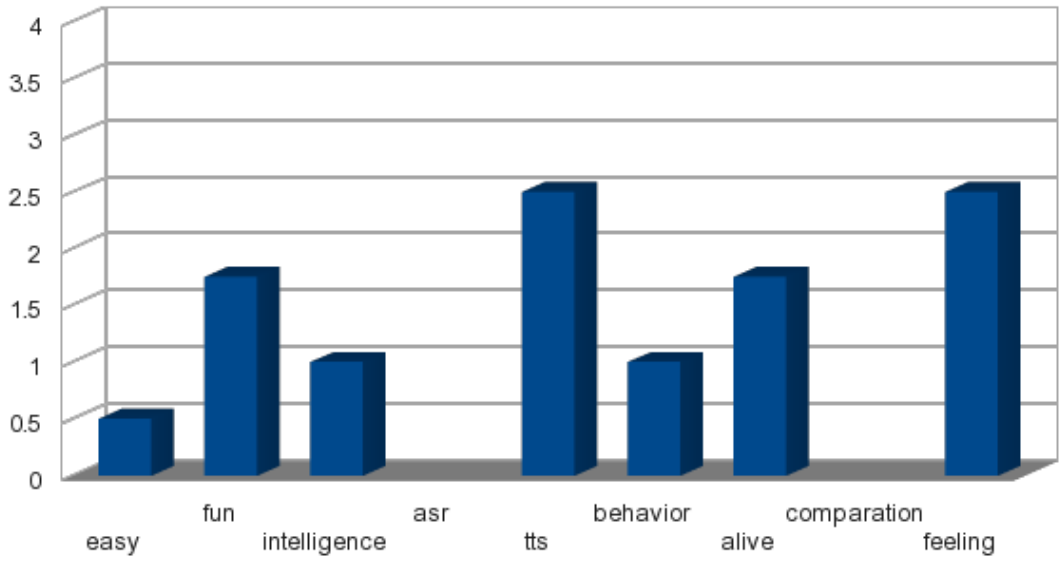


Figura A.8: Caso II. Después de una interacción de un grupo pequeño de niños inexpertos interactuando con Maggie. Cada columna representa el valor medio en el rango de 0-4, como el resultado de las preguntas sobre su subjetiva percepción de la interacción en los experimentos que han intervenido.

En la Figura A.8 se representa el valor medio de las respuestas a las mismas cuestiones que en el caso anterior. Como los usuarios sólo tenían 8 años, se ha adaptado las respuestas del cuestionario al rango de 0-4 (en lugar de 0 a 10 como en el caso I).

Como se esperaba, todos los chicos respondieron que el robot no era capaz de reconocer correctamente lo que ellos decían (el parámetro `asr`). Siendo este parámetro especialmente bajo, y como cabía esperar, es también muy bajo el valor medio del parámetro relacionado con la facilidad de uso, dado que se ve muy relacionado con la precisión del reconocimiento de voz.

Los usuarios pudieron entender lo que el robot les dijo (parámetro `etts`) pero tuvieron dificultades para entender su comportamiento (parámetro `comportamiento`). Por lo tanto en comparación con su ordenador personal o videoconsola, también este grupo de usuarios, seguían prefiriendo sus dispositivos electrónicos respecto al robot Maggie. Como era de experar, el parámetro `diversión` está por dejado del valor medio del rango dado (por debajo de 2), pero sin embargo los usuarios, como sensación general, se sintieron confortables con el uso del robot (parámetro `sensación producida`).

A.4.3. Caso III: interacción supervisada en pequeños grupos de niños

En este caso, los usuarios recibieron algunas instrucciones sobre como hablar al robot y los resultados reflejan un considerable mejora, respecto al caso anterior, de todos los parámetros analizados.

Los resultados son mostrados en la figura A.9. Un incremento en el parámetro `asr`, el único que mide como el usuario siente la capacidad del robot de entender el lenguaje hablado, implica un incremento importante en el resto de parámetros: `facilidad de uso`, `diversión` e `inteligencia`. El grado de felicidad y sensación general mientras se interactúa con el robot también se ve incrementado, aunque en menor medida, al tener menos rango de mejora.

Un incremento en la precisión del reconocimiento implica también un incremento en la percepción de la capacidad de entendimiento del robot, percibida por los usuarios. Esto puede ser explicado debido al hecho de que no sólo el usuario es entendido por el robot, sino que también el robot responde al usuario con las frases apropiadas, es decir con coherencia, por lo que los usuarios perciben un mejor comportamiento general.

A.4.4. Caso IV: interacción supervisada en grandes grupos de niños

En este caso, había mezclados en el experimento niños con experiencia y sin ella. Los resultados que se han obtenido son muy similar al caso anterior.

La principal diferente entre este caso y el caso III es el número de usuarios involucrados en los experimentos. Los resultados indican que el número de usuarios

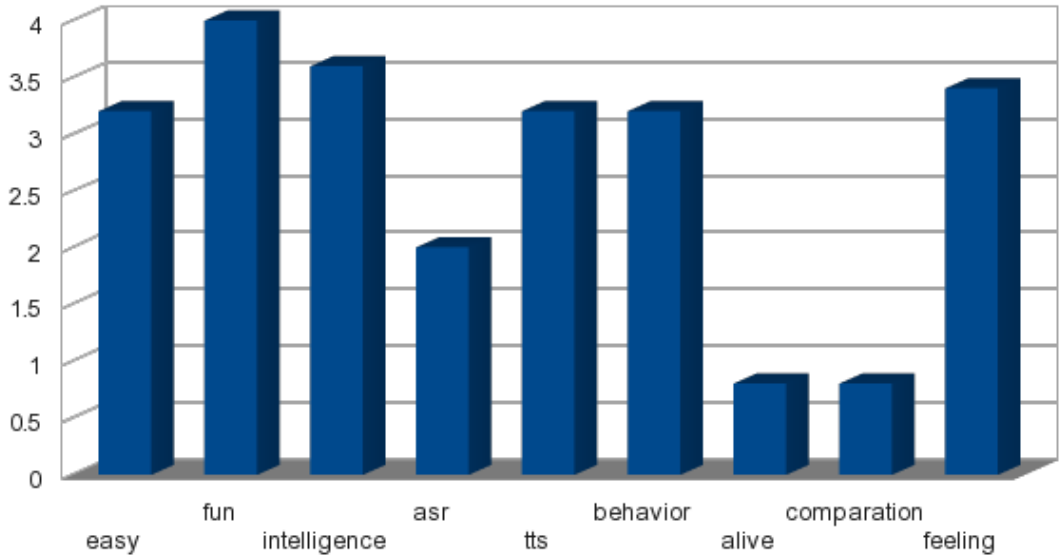


Figura A.9: Caso III. Valores para el grupo de chicos, con cierto conocimiento, interactuando con Maggie. Cada columna representa el valor medio para ese parámetro en el rango de 0 a 4.

involucrados en la interacción no es un gran obstáculo o impedimento para lograr entretenimiento con el robot, al igual que para el resto de parámetros, debido a que las métricas elegidas muestran resultados similares en ambos casos.

A.5. Conclusiones sobre los experimentos

Se ha probado la interacción entre jóvenes usuarios y Maggie, con su repertorio de habilidades y modos de interacción, en cuatro escenarios diferentes. Los experimentos con adolescentes muestran que el tiempo de interacción en el que se encuentran entretenidos es mucho menor que en usuarios más pequeños (niños de 8 años). Sin embargo no se ha podido medir cual es el tiempo máximo de interacción para el cual los usuarios se aburren de interactuar con el robot, ya que el tiempo que ha durado el experimento no ha sido suficiente para que cayesen en el hastío. En sus propias palabras, ellos pasarían todo el día jugando con Maggie.

La interacción verbal fue todavía lenta (4 turnos por minuto) comparada con la interacción humana, donde se suelen encontrar al menos 20 turnos por minuto; pero esta lentitud se ve en gran medida condicionada por la falta de experiencia de los

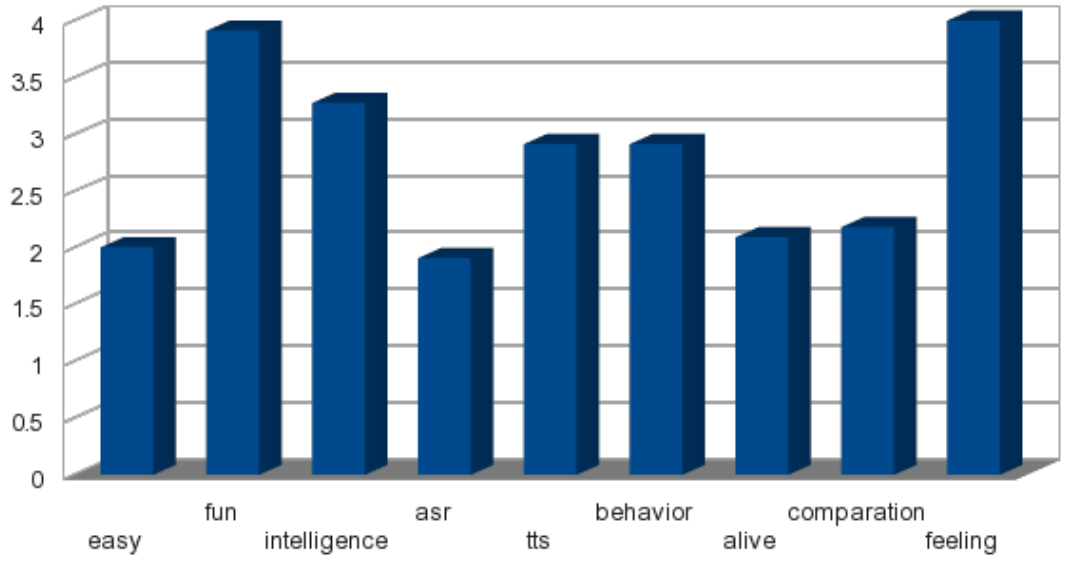


Figura A.10: Caso IV. Grupo grande de chicos interactuando con Maggie. Cada columna representa valores medios de entre 0 y 4, como resultado de las respuestas dadas a los cuestionarios al finalizar la interacción

usuarios con el robot, así como el propio modo de interacción, en modo tutorial, de Maggie cuando se encuentra con usuarios desconocidos para ella.

La precisión del reconocimiento de voz también fue inferior a la que se obtiene en interacciones naturales entre humanos con plenas capacidades auditivas. Algunas veces esto ocasiono un cuello de botella para el resto de las capacidades de interacción, lo que ha servido para potenciar y mejorar el desarrollo del resto de dichas capacidades. De está manera la interacción se vuelve menos dependiente del reconocimiento de voz. Un grado mayor de proactividad por parte del robot también es deseable.

En general, los usuarios inexpertos han necesitado algo de ayuda para interactuar con el robot: bien sea de otros usuarios con experiencia en la interacción o del propio grupo de desarrolladores. Sin embargo, la curva de aprendizaje ha sido muy rápida y los resultados indican que el sistema es muy fácil de usar una vez que se tiene cierto grado de experiencia con el mismo. Esto es debido, en parte, a las capacidades expresivas del robot y su facilidad de ser comprendidas.

APÉNDICE B

Gramáticas y diálogos de ejemplo

B.1. Ejemplos de gramáticas usadas

B.1.1. Gramática de bienvenida

```
#ABNF 1.0 ISO-8859-1;

language es-ES;
tag-format <semantics/1.0>;

public $root = [$GARBAGE] $language;

$language =(
   "hola magui" { out.option = "greet";  out.lang = "es"; } |
   "hola" { out.option = "greet";  out.lang = "es"; } |
   "hello maggie!en-gb" { out.option = "greet";  out.lang = "en"; } |
   "hello!en-gb" { out.option = "greet";  out.lang = "en"; } |
   "hi!en-gb" { out.option = "greet";  out.lang = "en"; } |
   "elou"{ out.option = "greet";  out.lang = "en"; } |
   "jelou"{ out.option = "greet";  out.lang = "en"; } |
   "jai"{ out.option = "greet";  out.lang = "en"; } |
   "jelou Magui"{ out.option = "greet";  out.lang = "en"; } |
   "jai Magui"{ out.option = "greet";  out.lang = "en"; } |
   "identificarme" { out.option = "login";  out.lang = "es"; } |
   "registrarme"   { out.option = "enroll"; out.lang = "es"; } |
   "salir"         { out.option = "exit";   out.lang = "es"; } |
   "loguin"   { out.option = "login";  out.lang = "en"; } |
   "identifai"   { out.option = "login";  out.lang = "en"; } |
   "enroll"  { out.option = "enroll"; out.lang = "en"; } |
   "anroll"  { out.option = "enroll"; out.lang = "en"; } |
   "sain ap"  { out.option = "enroll"; out.lang = "en"; } |
   "exit!en-gb"    { out.option = "exit";   out.lang = "en"; } |
   "identify!en-gb" { out.option = "login";  out.lang = "en"; } |
   "sign up!en-gb" { out.option = "enroll"; out.lang = "en"; } |
   "login|\\SAMPA=xsampa;(lA:g_}ln)"
        { out.option = "login";  out.lang = "en"; } |
   "enroll|\\SAMPA=xsampa;(Inr\oUl)|\\SAMPA=xsampa;(Enr\oUl)"
        { out.option = "enroll"; out.lang = "en"; } |
   "exit|\\SAMPA=xsampa;(Ek_hs@t_h)|\\SAMPA=xsampa;(Eg_}zlt_h)"
        { out.option = "exit";   out.lang = "en"; }
        );
```

B.1.2. Gramática Integrador

```

#ABNF 1.0 ISO-8859-1;

language es-ES;
tag-format <loq-semantics/1.0>;
public $root = $integrador;

$habilidad = (
        "adivinar personajes":AKINATOR |
        "habla mas bajo":BAJARVOLUMEN |
        "baterias":BATERIAS | "nivel de las  baterias":BATERIAS |
        "callate ya":CALLARSE |
        "no sigas diciendo mas cosas":CALLARSE |
        "cantame algo":CANTAR | "cantes":CANTAR | "canta":CANTAR |
        "quiero que cantes":CANTAR | "cantar":CANTAR |
        "sabes cantar":CANTAR |
```

```
17          "cuentame un chiste":CHISTE | "dime un chiste":CHISTE |
18          "contar caras":CONTARCARAS | "contar personas":CONTARCARAS |
19          "deportes":DEPORTES | "noticias deportivas":DEPORTES |
20          "correo electronico":EMAIL |
21          "escondite":ESCONDITE | "quien soy":ESCONDITE |
22          "estoy cansado de jugar contigo":FIN | "duermete":FIN |
23          "vete a dormir":FIN | "estoy harto de jugar contigo":FIN |
24          "me gustaria que te durmieras":FIN |
25          "me he cansado de jugar contigo":FIN |
26          "ya no quiero jugar mas contigo":FIN |
27          "ya no quiero jugar mas":FIN | "estoy harto":FIN |
28          "ahorcado":HANGMAN |
29          "que sabes hacer":HELP | "que puedes hacer":HELP |
30          "que opciones tienes:"HELP | "ayuda":HELP |
31          "no se que hacer:"HELP | "que hago":HELP |
32          "ayudame":HELP | "no se que hacer":HELP |
33          "que puedo hacer":HELP | "que le digo":HELP |
34          "me apetece jugar contigo":JUGAR |
35          "quiero jugar contigo":JUGAR |
36          "quiero jugar":JUGAR | "vamos a jugar":JUGAR |
37          "reconocer medicamentos":MEDICINAS |
38          "reconocer medicinas":MEDICINAS |
39          "deja de hacer esto":PARAR | "dejalo ya":PARAR |
40          "estate quieta":PARAR | "estate queto":PARAR | "detente":PARAR |
41          "para":PARAR | "otra cosa":PARAR | "parate":PARAR |
42          "callate un momento":PAUSESPEECH  |
43          "reconocer animales":PELUCHES | "reconocer peluches":PELUCHES |
44          "peluches":PELUCHES | "jugar con los peluches":PELUCHES |
45          "animales":PELUCHES | "jugar con los animales":PELUCHES |
46          "autopresentate":PRESENTACION | "presentate":PRESENTACION |
47          "ya puedes seguir hablando":RESUMESPEECH |
48          "persigueme":SIGUEME | "pilla pilla":SIGUEME | "seguir":SIGUEME |
49          "sigueme":SIGUEME | "vamonos de paseo":SIGUEME |
50          "vamos a dar una vuelta":SIGUEME | "vente conmigo":SIGUEME |
51          "habla mas alto":SUBIRVOLUMEN |
52          "conecta con la television":TELEMANDO |
53          "controla la television":TELEMANDO |
54          "maneja la television":TELEMANDO | "telemando":TELEMANDO |
55          "mover tu cuerpo":TELEOPERACION | "teleoperarte":TELEOPERACION |
56          "controlar tu cuerpo":TELEOPERACION |
57          "quiero controlar tu cuerpo":TELEOPERACION |
58          "quiero mover tu cuerpo":TELEOPERACION |
59          "quiero moverte":TELEOPERACION |
60          "tres en raya":TICTACTOE |
61          "informacion del tiempo":TIEMPO | "tiempo meteorologico":TIEMPO |
62          "tiempo para hoy":TIEMPO | "tiempo metereologico para hoy":TIEMPO |
63          "el tiempo de hoy":TIEMPO | "el tiempo":TIEMPO |
64          "tiempo para hoy":TIEMPO |
65          "volver al inicio":FIN |
66          "habla lo mas alto posible":VOLUMENMAX |
67          "habla lo mas bajo posible":VOLUMENMIN |
68          "quiero hacerte una pregunta":INFORMACION  |
69          "quiero hacerte unas preguntas":INFORMACION  |
70          "necesito informacion":INFORMACION |
71          "necesito saber una cosa":INFORMACION |
72          "necesito saber unas cosas":INFORMACION |
73          "quiero saber una cosa":INFORMACION |
74          "quiero esuchar musica":MUSIC | "ponme musica":MUSIC |
75          "me apetece escuchar musica":MUSIC |
76          "atacalo":ATACAR | "atacale":ATACAR | "ataca":ATACAR |
77          "como te llamas":ASKNAME | "dime tu nombre":ASKNAME |
```

```
78          "cual es tu edad":ASKAGE | "que edad tienes":ASKAGE |
79          "quiero que te despiertes":WAKEUP | "despiertate":WAKEUP
80
81  ){<@Skill $value>};
82
83  $integrador = [$GARBAGE] $habilidad;
```

B.1.3. Gramática para la teleoperación multimodal del robot

```
1
2   #ABNF 1.0 ISO-8859-1;
3
4   language es-ES;
5   tag-format <semantics/1.0>;
6
7   public $root = [$GARBAGE] $action;
8
9   $action =(
10          "subelo" { out.Action = "subir"; } |
11          "bajalo" { out.Action = "bajar"; } |
12          "sube el brazo derecho" { out.Action = "subir"; out.BodyPart = "brazo_derecho
                "; } |
13          "sube el brazo izquierdo" { out.Action = "subir"; out.BodyPart = "
                brazo_izquierdo"; } |
14          "baja el brazo derecho" { out.Action = "bajar"; out.BodyPart = "brazo_derecho
                "; } |
15          "baja el brazo izquierdo" { out.Action = "bajar"; out.BodyPart = "
                brazo_izquierdo"; } |
16          "girala a la derecha" { out.Action = "girarderecha"; } |
17          "girala a la izquierda" { out.Action = "girarizquierda"; } |
18          "gira la cabeza a la derecha" { out.Action = "girarderecha"; out.BodyPart = "
                cabeza"; } |
19          "gira la cabeza a la izquierda" { out.Action = "girarizquierda"; out.BodyPart
                 = "cabeza"; } |
20          "vete a la derecha" { out.Action = "moverderecha"; out.BodyPart = "base"; } |
21          "vete a la izquierda" { out.Action = "moverizquierda"; out.BodyPart = "base";
                 } |
22          "retrocede" { out.Action = "retroceder"; out.BodyPart = "base"; } |
23          "avanza" { out.Action = "avanzar"; out.BodyPart = "base"; } |
24          "estate quieta" { out.Action = "parar"; } |
25          "parate" { out.Action = "parar"; } |
26          "para" { out.Action = "parar"; } |
27          "dejalo ya" { out.Action = "parar"; } |
28          "no lo hagas mas" { out.Action = "parar"; } |
29          "vete al cargador" {out.Action = "cargador"; out.BodyPart = "base";} |
30          "acercate al caragdor" {out.Action = "cargador"; out.BodyPart = "base";} |
31          "vete a la puerta" {out.Action = "puerta"; out.BodyPart = "base";} |
32          "acercate a la puerta" {out.Action = "puerta"; out.BodyPart = "base";} |
33          "acercate a la television" {out.Action = "television"; out.BodyPart = "base"
                ;} |
34          "vete a la television" {out.Action = "television"; out.BodyPart = "base";}
35  );
```

B.1.4. Gramática para la selección de juegos

```
1   #ABNF 1.0 ISO-8859-1;
2
3   language es-ES;
4   tag-format <loq-semantics/1.0>;
```

```
5  public $root = $integrador;
6
7  $juego = (
8      "adivinar personajes":AKINATOR |
9      "habla mas bajo":BAJARVOLUMEN |
10     "callate ya":CALLARSE |
11     "no sigas diciendo mas cosas":CALLARSE |
12     "escondite":ESCONDITE | "quien soy":ESCONDITE |
13     "ahorcado":HANGMAN |
14     "deja de hacer esto":PARAR | "dejalo ya":PARAR | "estate quieta":PARAR |
15     "callate un momento":PAUSESPEECH |
16     "reconocer animales":PELUCHES | "reconocer peluches":PELUCHES |
17     "peluches":PELUCHES | "jugar con los peluches":PELUCHES |
18     "animales":PELUCHES | "jugar con los animales":PELUCHES |
19     "ya puedes seguir hablando":RESUMESPEECH |
20     "persigueme":SIGUEME | "pilla pilla":SIGUEME | "seguir":SIGUEME |
21     "sigueme":SIGUEME | "vamonos de paseo":SIGUEME |
22     "vamos a dar una vuelta":SIGUEME | "vente conmigo":SIGUEME |
23     "habla mas alto":SUBIRVOLUMEN |
24     "tres en raya":TICTACTOE |
25     "habla lo mas alto posible":VOLUMENMAX |
26     "habla lo mas bajo posible":VOLUMENMIN
27 ){<@Game $value>};
28
29 $integrador = [$GARBAGE] $juego;
```

B.1.5. Gramática para el control de la televisión

```
1  #ABNF 1.0 ISO-8859-1;
2
3  language es-ES;
4  tag-format <loq-semantics/1.0>;
5
6  public $root = [$GARBAGE] $telemando;
7
8  $telemando = (
9          "enciende la television":ON | "enciende la tele":ON |
10         "apaga la tele":OFF | "apaga la television":OFF |
11         "sube un canal":UP | "baja un canal":DOWN |
12         "pon la radio":FM | "pon la television digital terrestre":TDT |
13         "sube el volumen":VOLUMENMAS | "pon la tele":TDT | "pon la television":TDT |
14         "baja el volumen":VOLUMENMENOS | "deja de manejar la television":FINTV |
15         "deja de controlar la television":FINTV | "deja de manejar la tele":FINTV |
16         "deja de controlar la tele":FINTV | "deja la tele":FINTV |
17         "deja la television":FINTV
18 ){<@Comando $value>};
```

B.2. Diálogos de voz siguiendo el estándar VoiceXML

B.2.1. Diálogo de entrada al sistema

```
1  <?xml version="1.0" encoding="ISO-8859-1"?>
2  <vxml xmlns="http://www.w3.org/2001/vxml"
3        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
4        xsi:schemaLocation="http://www.w3.org/2001/vxml http://www.w3.org/TR/voicexml20
             /vxml.xsd"
```

```
5       version="2.0"
6       application="rootDocument.vxml"
7      >
8
9
10
11
12 <form id="Main">
13
14     <property name="timeout" value="5s"/>
15      <var name = "n" expr = "0"/>
16
17
18      <script><![CDATA[
19      // Script para obtener mediante data valores guardados en documentos XML
            independientes
20      //retrieve the value contained in the node t from the DOM exposed by d
21      function GetData(documentoXML, nombreUsuario,  propiedadObtener){
22         try{
23
24            var usuariosNodeList = documentoXML.getElementsByTagName("user");
25
26            for( var i = 0; i < usuariosNodeList.length; i++ ){
27                 var usuarioNode = usuariosNodeList.item(i);
28                 if (usuarioNode.hasAttributes() == true){
29                     if (usuarioNode.hasAttribute('name') == true){
30                         var nameUser  = usuarioNode.getAttribute('name');
31
32                         if (usuarioNode.hasAttribute(propiedadObtener) == true ){
33
34                            if ((nameUser+"" == nombreUsuario)){
35                               var valorDevolver = usuarioNode.getAttribute(
                                     propiedadObtener);
36                               return valorDevolver+"";
37                           }
38                         }
39                     }
40
41                 }
42             }
43             return -2;
44         }catch(e){
45             return -1;  // the value could not be retrieved, so return this instead
46         }
47      }
48
49
50      // Script para obtener mediante data valores guardados en documentos XML
            independientes
51      //retrieve the value contained in the node t from the DOM exposed by d
52      function SetData(documentoXML, nombreUsuario,  propiedadCambiar){
53         //try{
54
55            var usuariosNodeList = documentoXML.getElementsByTagName("user");
56
57            for( var i = 0; i < usuariosNodeList.length; i++ ){
58                 var usuarioNode = usuariosNodeList.item(i);
59                 if (usuarioNode.hasAttributes() == true){
60                     if (usuarioNode.hasAttribute('name') == true){
61                         var nameUser  = usuarioNode.getAttribute('name');
62
```

```
63                          if (usuarioNode.hasAttribute(propiedadCambiar) == true ){
64
65                              if ((nameUser+"" == nombreUsuario)){
66                                  var experience = usuarioNode.getAttribute(
                                        propiedadCambiar);
67                                  var entero = parseInt(experience)+1;
68                                  usuarioNode.setAttribute('experience',entero);
69                                  return entero;
70                              }
71                          }
72                      }
73
74                  }
75              }
76              return -2;
77          /*}catch(e){
78              return -1;  // the value could not be retrieved, so return this instead
79          }*/
80      }
81      ]]></script>
82
83      <block>
84                  <!-- Spanish Conversation -->
85
86                  <!-- Non verbal communication -->
87                  <prompt>\emotion=happy</prompt>
88                  <prompt>#emit$NEW_STATE_REQUEST$2</prompt>
89                  <prompt>#emit$EMOTION_SPEECH_DETECTION_SKILL_START</prompt>  <!--
                        Activamos la deteccion de emociones -->
90                  <!--<prompt>#emit$GESTURE_STOP_LOOP</prompt>
91                  <prompt>#emit$GESTURE_ZERO</prompt>-->
92                  <prompt>#emit$PARPADEO_INIT</prompt>
93                  <prompt>#emit$VOICE_TRACKER_SKILL_START</prompt>
94                  <prompt>#Sound$HELLO</prompt>
95                  <prompt>#Sound$HELLO</prompt>
96                  <prompt>#emit$GESTURE_STOP_LOOP</prompt>
97                  <prompt>#emit$GESTURE_TRALARA</prompt>
98                  <prompt count="1">#setGrammar$enrollOrLogin.gram</prompt>
99                  <prompt>#multiverifcation</prompt>
100         </block>
101         <grammar src = "../../asr/Grammars/trivial.grxml" type="application/srgs+xml"
                mode="voice" xml:lang="es-ES" version="1.0"/>
102
103
104         <!-- Para fijar la variable lenguaje que sera comun a todos los documentos,
                ya que esta declarada en el rootDocument.xml -->
105         <field name = "lang">
106             <filled>
107                 <!--prompt>#emit$GESTURE_GAZE</prompt-->
108                 <prompt>#emit$GESTURE_HEAD_ASSERT$1</prompt>
109                 <if cond = "lang == 'es'">
110                     <prompt>#setLanguage$es</prompt>
111                     <assign name = "language" expr= "'es'"/>
112                     <assign name = "application.language" expr= "'es'"/>
113                 <elseif cond = "lang == 'en'" />
114                     <prompt>#setLanguage$en</prompt>
115                     <assign name = "language" expr= "'en'"/>
116                     <assign name = "application.language" expr= "'en'"/>
117                 </if>
118             </filled>
119         </field>
```

```
120
121         <field name="option">
122             <filled>
123                         <prompt>#emit$NEW_STATE_REQUEST$2</prompt>
124                         <prompt>#emit$ESPERANDO_START</prompt>
125
126                         <if cond = "option == 'login'">
127                                 <goto next = "login.vxml" />
128                         <elseif cond = "option == 'enroll'"/>
129                                 <goto next = "enroll.vxml" />
130                         <elseif cond = "option == 'greet'"/>
131
132                           <if cond = " lastresult$.nameuserspeaking == 'unknown' ">
133                                <!-- <prompt>es:\emotion=sad</prompt> -->
134                                <prompt>es:No te conozco.</prompt>
135                                <prompt>#emit$GESTURE_HEAD_DENY$1</prompt>
136                                <prompt>es:\emotion=happy</prompt>
137                                <goto next = "#SubD_Response" />
138
139                           <else/>  <!-- USUARIO IDENTIFICADO -->
140                               <if cond = "lastresult$.scoresi &lt; 2.5">
141                                   <prompt>#emit$GESTURE_HEAD_DENY$1</prompt>
142                                   <assign name = "n" expr = "n +1 "/>
143                                  <if cond="n &lt; 3">
144                                     <prompt>es: Creo que eres <value expr="
                                            lastresult$.nameuserspeaking"/>. Pero no
                                            estoy seguro. Vuelve a saludarme. </prompt>
145                                     <goto nextitem = "lang" />
146                                  <else/>
147
148                                     <prompt>es:No te conozco</prompt>
149                                     <prompt>es:\emotion=happy</prompt>
150                                     <goto next = "#SubD_Response" />
151                                   </if>
152
153                               <else/>
154
155                                 <prompt>#sound$AMAZING</prompt>
156                                 <prompt>#emit$GESTURE_YUPI$1</prompt>
157                                 <!--<prompt>#NLG$HELLO</prompt> -->
158                                 <prompt>\NLG=HELLO <value expr="lastresult$.
                                        nameuserspeaking"/></prompt>
159                                 <assign name = "application.userName" expr= "
                                        lastresult$.nameuserspeaking"/>
160
161
162
163                                 <!-- ____________________CARGAMOS EL PERFIL DEL
                                        USUARIO_________________ -->
164                                 <!-- Abrimos el fichero con la informacion -->
165
166                                 <var name="directorio" />
167                                 <if cond = "language == 'es'">
168                                    <assign name ="directorio" expr="'esES'" />
169                                 <elseif cond = "language == 'en'"/>
170                                     <assign name ="directorio" expr="'enGB'" />
171                                 </if>
172                                 <data name="users" srcexpr = " '../../asr/
                                        KnowledgeBase/' +  directorio + '/users.xml'" />
```

```
173                                 <!--<data name="users" src = "file:/home/user/manager
                                        /long_term_memory/asr/KnowledgeBase/esES/users.
                                        xml"/> -->
174
175                                 <!-- Incrementamos la experiencia -->
176                                 <prompt>#addExperience$<value expr="language"/>$<
                                        value expr="application.userName" /></prompt>
177 <!--
178                                 <var name="valorDevuelto" expr="SetData(users,
                                        application.userName, 'experience')"/>
179                                 <prompt>es: Nuevo valor de experiencia <value expr="
                                        valorDevuelto"/> </prompt>
180 -->
181
182                                 <!-- Cargamos la edad -->
183                                 <var name="valorDevuelto" expr="GetData(users,
                                        application.userName, 'age')"/>
184                                 <prompt>es: Tienes una edad de <value expr="
                                        valorDevuelto"/> </prompt>
185                                 <assign name = "application.userAge" expr= "
                                        valorDevuelto"/>
186
187                                 <!-- Cargamos la experiencia -->
188
189                                 <var name="valorDevuelto" expr="GetData(users,
                                        application.userName, 'experience')"/>
190                                 <prompt>es: Tienes una experiencia de <value expr="
                                        valorDevuelto"/> </prompt>
191                                 <assign name = "application.experience" expr= "
                                        valorDevuelto"/>
192
193
194                                 <!-- Vamos al integrador -->
195                                 <goto next = "integrator.vxml" />
196                               </if>
197                           </if>
198                         <elseif cond = "option == 'rfidGreet'" />
199                             <prompt>#sound$AMAZING</prompt>
200                             <prompt>#emit$GESTURE_YUPI$1</prompt>
201                             <goto next = "integrator.vxml"/>
202
203                         <elseif cond = "option == 'exit'"/>
204                                 <prompt>es:Adios. Hasta otra.</prompt>
205                                 <exit/>
206                         </if>
207         </filled>
208
209         <nomatch>
210                         <prompt>#Sound$ERROR</prompt>
211                         <prompt>#emit$GESTURE_HEAD_DENY$1</prompt>
212                         <goto nextitem = "option" />
213         </nomatch>
214
215         <noinput>
216             <!-- <prompt>#NLG$NOT_INPUT</prompt> -->
217             <prompt>es:Hola</prompt>
218             <goto nextitem = "option"/>
219         </noinput>
220
221
222
```

```
223     </field>
224 </form>
225
226 <form id="SubD_Response">
227     <subdialog name="SubD_SiONo" src="response_yes_no.vxml">
228
229         <param name="myquestion"  expr="Quieres responderme a unas preguntas para
                conocernos mejor?"  />
230         <filled>
231                 <if cond="SubD_SiONo.Respuesta == 'SI'">
232                         <prompt>#emit$GESTURE_HEAD_ASSERT$1</prompt>
233                         <goto next = "enroll.vxml" />
234                 <else/>
235                          <prompt>#emit$GESTURE_HEAD_DENY$1</prompt>
236                          <prompt>Vale. No. </prompt>
237                          <goto next = "integrator.vxml" />
238                 </if>
239         </filled>
240    </subdialog>
241 </form>
242 </vxml>
```

B.2.2. Diálogo de registro en el sistema

```
1  <?xml version="1.0" encoding="ISO-8859-1"?>
2  <vxml xmlns="http://www.w3.org/2001/vxml"
3        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
4        xsi:schemaLocation="http://www.w3.org/2001/vxml http://www.w3.org/TR/voicexml20
             /vxml.xsd"
5        version="2.0"
6        application="rootDocument.vxml"
7       >
8
9  <!-- Para almacenar informacion de la voicePrint -->
10 <var name = "name"/>
11 <var name = "age"/>
12
13 <form id="Enroll">
14
15         <grammar src = "../../asr/Grammars/trivial.grxml" type="application/srgs+xml"
               mode="voice" xml:lang="es-ES" version="1.0"/>
16
17         <var name = "n" expr = "1"/>
18
19         <field name="Idioma">
20                 <!-- Fijamos una unica gramatica multilenguaje -->
21                 <prompt count="1">#withoutMultiverification</prompt>
22         <prompt count="1">#setGrammar$idiomas.gram</prompt>
23                 <prompt>
24                 es:Cual es tu idioma preferido?|en:What is your favourite language?
                       </prompt>
25
26                 <filled>
27                         <prompt>#emit$GESTURE_HEAD_ASSERT$1</prompt>
28                         <if cond = "Idioma == 'ENGLISH'">
29                                 <prompt>#setLanguage$en</prompt>
30                                 <prompt>Perfect. I know that you prefer a English
                                       conversation </prompt>
31                                 <assign name = "application.language" expr= "'en'"/>
32                         <elseif cond = "Idioma == 'SPANISH'" />
```

```
33                              <assign name = "application.language" expr = "'es'"/>
34                              <prompt>#setLanguage$es</prompt>
35                              <prompt>#emit$GESTURE_WINK_LEFT</prompt>
36                              <prompt>Hablaremos tu y yo siempre en castellano </
                                    prompt>
37                      </if>
38                      <assign name = "n" expr = "1"/>
39              </filled>
40
41              <noinput>
42                      <prompt>es:Dime tu idioma|en:Your language.Please. </prompt>
43                      <assign name = "n" expr = "n +1 "/>
44              </noinput>
45      </field>
46
47
48      <field name="Name">
49              <prompt cond = "language == 'en'">#setGrammar$enGB/names.gram</prompt
                    >
50              <prompt cond = "language == 'es'">#setGrammar$nombres.gram</prompt>
51              <prompt>es:Me gustaria saber como te llamas. Dime tu nombre por favor
                    ..|en:I need know your name. Sorry. Say me your name."</prompt>
52
53              <filled>
54                      <prompt>Hola <value expr = "lastresult$.utterance"/></prompt>
55                      <assign name = "name" expr = "lastresult$.utterance"/>
56                      <assign name = "application.userName" expr= "lastresult$.
                            utterance"/>
57                      <prompt>#emit$GESTURE_GREET$1</prompt>
58              <assign name = "n" expr = "1"/>
59              </filled>
60
61              <noinput>
62                      <prompt>es:Dime tu nombre.|en:your name. Please. </prompt>
63                      <assign name = "n" expr = "n +1 "/>
64              </noinput>
65
66      </field>
67
68
69      <field name="Entero">
70
71              <prompt cond = "language == 'en'">#setGrammar$enGB/integer.gram</
                    prompt>
72              <prompt cond = "language == 'es'">#setGrammar$integer.gram</prompt>
73              <prompt>es:Cuantos anyos tienes?|en:How old are you?</prompt>
74
75              <filled>
76                      <prompt>#emit$GESTURE_STOP_LOOP</prompt>
77                      <prompt>#emit$GESTURE_HEAD_ASSERT$1</prompt>
78                      <prompt><value expr = "lastresult$.utterance"/></prompt>
79                      <assign name = "age" expr = "Entero"/>
80                      <assign name = "application.userAge" expr= "Entero"/>
81                      <prompt>es:Ahora quiero aprenderme tu tuno de voz. Saludame.
                            </prompt>
82                      <goto next = "#InvokeSubDialgo" />
83              </filled>
84
85              <noinput>
86                      <prompt>es:Dime tu edad|en:your age. Please </prompt>
87                      <assign name = "n" expr = "n +1 "/>
```

```
88                  </noinput>
89
90          </field>
91
92  </form>
93
94  <!-- Formulario desde el que vamos a invocar al subdialogo que nos va a registrar las
        huellas de voz -->
95  <form id="InvokeSubDialgo">
96          <subdialog name="Voiceprints" src="voiceprints.vxml">
97                  <param name="name" expr="document.name" />
98                  <param name="age" expr="document.age" />
99                  <param name="language" expr="application.language" />
100
101                 <filled>
102                         <prompt>#emit$GESTURE_STOP_LOOP$1</prompt>
103                         <goto next = "integrator.vxml"/>
104                 </filled>
105
106         </subdialog>
107
108 </form>
109
110 </vxml>
```

APÉNDICE C

Glosario de Acrónimos

- AC (Acto Comunicativo): unidad mínima de intercambio de información mediante diálogo natural. En la práctica, para nuestro sistema, está información es formalizada mediante el estándar NLSML e incluye toda la información multimodal obtenida durante un intervalo de tiempo. Es consecuencia de la adaptación de la teoría de actos comunicativos en la interacción entre humanos a nuestro sistema artificial de interacción por diálogos. Los actos comunicativos pueden ser expresivos (salida), si los realiza uno mismo, o de percepción (entrada) si los realiza el otro interlocutor.
- API (Application Programming Interface): es el conjunto de funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción y comunicación.
- ASR (Automatic Speech Recognition): Reconocimiento Automático del Habla mediante computador. Es la capacidad de transcribir voz a texto escrito.
- AU (Action Unit): es la unidad mínima de la que se forman las expresiones faciales. Cada una de estas AUs tiene distintos niveles de intensidad. Forma parte de la teoría FACS.
- BML (Behaviour Markup Language): lenguaje en XML que especifica como codificar actos comunicativos de expresión multimodal para robot humanoides. En este lenguaje se especifican marcas que permite la sincronización temporal entre los diversos modos.

- DMS (Dialog Manager System): Es el módulo del sistema de diálogo que se encarga de la gestión del diálogo, es decir de teniendo en cuenta las entradas que generan los módulos de percepión, genera movimientos en el flujo del diálogo y acciones (actos comunicativos de expresión).

- ECMAScript: Lenguaje de javascript incrustado dentro de las gramáticas para definir las reglas semánticas de la gramática.

- ETTS (Emotional Text To Spech): Síntesis de texto a voz con recreación artificial de emociones.

- FACS (Facial Action Coding System) sistema que codifica las expresiones faciales como un conjunto de unidades básicas (AUs) susceptibles de ser combinadas para expresar emociones en el rostro.

- Gramáticas: Conjunto de reglas que establecen las combinaciones de palabras válidas para un subconjunto del lenguaje. Si además añaden reglas semánticas, se conocen como gramáticas semánticas.

- VoiceXML: lenguaje estándar de especificación de aplicaciones de voz que facilita su desarrollo. Se basa en el paradigma de rellenado de huecos de información. Existen varias versiones de este estándar. Partiendo del estándar se pueden hacer extensiones o modificaciones para adaptar su uso, tal y como se ha hecho en este trabajo.

- HRI (Human-Robot Interaction): Es el tipo de interacción que se produce entre uno o varios humanos, con uno o varios robots. Constituye el campo de investigación fundamental de este trabajo.

- IBL (Instruction Based Learning): aprendizaje de robots por instrucciones.

- IDiM (Interpretad Dialog Manager): es el nombre del gestor de diálogo del sistema de diálogo RDS aquí presentado.

- NLG (Natural Language Generation): es el conjunto de técnicas para generación de lenguaje natural, es decir de sintetizar voz partiendo de conceptos abstractos.

- NLP (Natural Language Programming): programación de acciones, comportamientos o secuencias en base a interacción natural por diálogos, sin necesidad de pre-programarlas usando lenguajes no naturales, como los de programación (C, Java, etc). Es una manera mucho más humana de aprendizaje que mediante lenguaje de programación.

- NLP (Natural Language Processing): es el conjunto de tecnologías que engloban NLG y NLU.
- NLSML (Natural Language Semantics Markup Language): Lenguaje estándar basado en XML utilizado para representar los valores semánticos asociados a cada acto comunicativo después de la fusión multimodal.
- NLU (Natural Language Understanding): es el conjunto de técnicas que permiten extraer significado relevante para el diálogo fruto del análisis del lenguaje natural. Dentro de este campo se encuentra el ASR.
- POMDP (Partially Observable Markov Decision Process): es una generalización de un MPD, en el que se tiene cierta incertidumbre sobre las entradas de información del sistema. En el caso de un sistema para dialogar la incertidumbre suele estar dada por el reconocimiento de voz. Normalmente un sistema basado en un POMDP tiene un coste computacional muy elevado.
- RDS (Robotic Dialog System): Sistema de Diálogo Robótico implementado para esta tesis. Conforma un conjunto de módulos de entrada y salida de información, así como módulos para su gestión. Pretende convertirse en un estándar en la HRI.
- Robot Social: Subconjunto de los robots. Están diseñados para interactuar de *manera natural* con los humanos, cumpliendo tareas como entretenimiento, asistencial, etc.
- SRGS (Speech Recognition Grammar Specification): Estándar que define como implementar gramáticas válidas para cualquier reconocedor de voz.
- SVM (Support Vector Machine): técnica de aprendizaje automático que tomando en cuenta una entrada dada puede clasificarla en un conjunto de entre los posibles. Para la construcción de el modelo que clasifica es necesario un conjunto de datos de entrenamiento.

Bibliografía

[A.Corrales, R.Rivas, 2009] A.Corrales, R.Rivas, M. (2009). Integration of a RFID System in a social robot. *Computer and Information Science*, 44(10.1007/978-3-642-03986-7_8), 63–73.

[Allen, 1999] Allen, J. F. (1999). Mixed-initiative interaction. *IEEE Intelligent Systems*, 14(5).

[Allen et al., 2001] Allen, J. F., Byron, D. K., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Towards conversational human-computer interaction. *AI Magazine*, 22(4), 27–38.

[Alonso & Salichs, 2011] Alonso, F. & Salichs, M. A. (2011). Integration of a voice recognition system in a social robot. *Cybernetics and Systems: An International Journal*, 42(4), 215–245.

[Alonso-Martin et al., 2013] Alonso-Martin, F., Castro-González, A., Gorostiza, J., & Salichs, M. A. (2013). Multidomain Voice Activity Detection during Human-Robot Interaction. In *International Conference on Social Robotics (ICSR 2013)* (pp. 64–73). Bristol: Springer International Publishing.

[Alonso-Martín et al., 2012] Alonso-Martín, F., Gorostiza, J., Malfaz, M., & Salichs, M. (2012). User Localization During Human-Robot Interaction. *Sensors*, 12(7), 9913–9935.

[Alonso-Martín et al., 2013] Alonso-Martín, F., Gorostiza, J. F., Malfaz, M., & Salichs, M. (2013). Multimodal Fusion as Communicative Acts during Human-Robot Interaction. *Cybernetics and Systems*, 44(8), 681–703.

[Alonso-Martin et al., 2012] Alonso-Martin, F., Gorostiza, J. F., & Salichs, M. A. (2012). Musical Expression in a Social Robot. In *Proceedings of the 2012 International IEEE Intelligent Vehicles Symposium. Workshops V Perception in Robotics.* (pp. 45–59). Alcalá de Henares.

[Alonso-Martin et al., 2013a] Alonso-Martin, F., Gorostiza, J. F., & Salichs, M. A. (2013a). Preliminary Experiments on HRI for improvement the Robotic Dialog System (RDS). In *Robocity2030 11th Workshop: Robots Sociales* Leganés (Spain).

[Alonso-Martin et al., 2013b] Alonso-Martin, F., Malfaz, M., Sequeira, J., Gorostiza, J., & Salichs, M. A. (2013b). A Multimodal Emotion Detection System during Human-Robot Interaction. *Sensors*, 13(11), 15549–15581.

[Alonso-Martin et al., 2011] Alonso-Martin, F., Ramey, A. A., & Salichs, M. A. (2011). Maggie: el robot traductor. In UPM (Ed.), *9º Workshop RoboCity2030-II*, number Breazeal 2003 (pp. 57–73). Madrid: Robocity 2030.

[Alonso-Martin & Salichs, 2011] Alonso-Martin, F. & Salichs, M. (2011). Integration of a voice recognition system in a social robot. *Cybernetics and Systems*, 42(4), 215–245.

[Andersson et al., 2004] Andersson, S., Handzel, A., Shah, V., & Krishnaprasad, P. (2004). Robot phonotaxis with dynamic sound-source localization. In *Robotics and Automation, 2004. Proceedings. ICRA'04. April*, volume 5 (pp. 4833–4838). New Orleans (EEUU): IEEE.

[Argyle, 1988] Argyle, M. (1988). *Bodily communication.* New York, NY, US: Methuen.

[Argyle, M.; Dean, 1965] Argyle, M.; Dean, J. (1965). Eye-contact, distance and affiliation. In *Sociometry.*

[Arnold, 1960] Arnold, M. (1960). *Emotion and personality.* Columbia University Press.

[Bach & Harnish, 1979] Bach, K. & Harnish, R. (1979). *Linguistic communication and speech acts.* Cambridge.

[Barber & Salichs, 2001a] Barber, R. & Salichs, M. (2001a). A new human based architecture for intelligent autonomous robots. In *4th IFAC Symposium on Intelligent Autonomous Vehicles* (pp. 85–90). Sapporo (Japan): Pergamon.

[Barber & Salichs, 2001b] Barber, R. & Salichs, M. (2001b). A new human based architecture for intelligent autonomous robots. In *The Fourth IFAC Symposium on Intelligent Autonomous Vehicles* (pp. 85–89).

[Barrett, 1998] Barrett, P. (1998). Voice activity detector. *US Patent 5,749,067*.

[Barrett, 2000] Barrett, P. (2000). Voice activity detector. *US Patent 6,061,647*.

[Bartlett et al., 2003] Bartlett, M. S., Littlewort, G., Fasel, I., & Movellan, J. R. (2003). Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. In *2003 Conference on Computer Vision and Pattern Recognition Workshop* (pp. 53–53).: IEEE.

[Bassili, 1978] Bassili, J. (1978). Facial motion in the perception of faces and of emotional expression. *Journal of Experimental Psychology: Human Perception and Performance, 4*, (pp. 373–379).

[Benesty et al., 2008] Benesty, J., Chen, J., & Huang, Y. (2008). *Microphone array signal processing*, volume 1. Springer Verlag.

[Bennett et al., 2002] Bennett, C., Llitjós, A., Shriver, S., Rudnicky, A., & Black, A. W. (2002). Building VoiceXML-based applications. In *Seventh International Conference on Spoken Language Processing* (pp. 2–5).: Citeseer.

[Billard et al., 2007] Billard, A., Robins, B., Nadel, J., & Dautenhahn, K. (2007). Building Robota, a mini-humanoid robot for the rehabilitation of children with autism. *Assistive technology : the official journal of RESNA*, 19(1), 37–49.

[Birdwhistell, 1970] Birdwhistell, R. L. (1970). *Kinesics and Context. Essays on Body Motion Communication*. Philadelphia, USA: University of Pennsylvania Press.

[Bischoff & Graefe, 2004] Bischoff, R. & Graefe, V. (2004). HERMES - a versatile personal robotic assistant. *Proceedings of the IEEE*, 92(11), 1759–1779.

[Boersma, 2002] Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10), 341–345.

[Bohus & Horvitz, 2009] Bohus, D. & Horvitz (2009). Dialog in the open world: Platform and applications. In *Proceedings of ICMI'09* MA, USA.

[Bohus & Horvitz, 2010] Bohus, D. & Horvitz, E. (2010). On the challenges and opportunities of physically situated dialog. In *Dialog with Robots AAAI 2010 Fall Symposium* Virgina, USA.

[Bohus et al., 2007] Bohus, D., Raux, A., Harris, T. K., Eskenazi, M., & Rudnicky, A. I. (2007). Olympus: an open-source framework for conversational spoken language interface research. (pp. 32–39).

[Bohus & Rudnicky, 2009] Bohus, D. & Rudnicky, A. (2009). The ravenclaw dialog management framework: Architecture and systems. *Computer Speech and Language*, 23(3), 332 – 361.

[Bos et al., 2003] Bos, J., Klein, E., Lemon, O., & Oka, T. (2003). Dipper: Description and formalisation of an information-state update dialogue system architecture. In *4th SIGdial Workshop on Discourse and Dialogue* Sapporo, Japan.

[Bos & Oka, 2003] Bos, J. & Oka, T. (2003). Building spoken dialogue systems for believable characters. In *4th SIGdial Workshop on Discourse and Dialogue* Sapporo, Japan.

[Bradley & Lang, 1994] Bradley, M. & Lang, P. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry, 25*.

[Bradley & Lang, 2000] Bradley, M. & Lang, P. (2000). Measuring emotion: Behavior, feeling, and physiology. In *Cognitive neuroscience of emotion* (pp. 242–276). New York, NY, US: Oxford University Press.

[Brandstein & Ward, 2001] Brandstein, M. & Ward, D. (2001). *Microphone arrays: signal processing techniques and applications*. Springer Verlag.

[Breazeal, 2001] Breazeal, C. (2001). Emotive Qualities in Robot Speech Approved for Public Release. *Artificial Intelligence*, (1993), 1–7.

[Breazeal et al., 2004] Breazeal, C., Brooks, A., Chilongo, D., Gray, J., Hoffman, A., Lee, C. K. H., Lieberman, J., & Lockered, A. (2004). Working collaboratively with humanoid robots. *ACM Computers in Entertainment*, 2(3).

[Breazeal et al., 2003] Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lieberman, J., Lockerd, A., & Mulanda, D. (2003). Humanoid Robots as Cooperative Partners for People. *International Journal*.

[Briere et al., 2008] Briere, S., Valin, J.-M., Michaud, F., & Letourneau, D. (2008). Embedded auditory system for small mobile robots. In *2008 IEEE International Conference on Robotics and Automation* (pp. 3463–3468).: IEEE.

[Bronkhorst, 2000] Bronkhorst, A. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*.

[Brown, 1874] Brown, A. C. (1874). The Sense of Rotation and the Anatomy and Physiology of the Semicircular Canals of the Internal Ear. *Journal of anatomy and physiology*, 8(Pt 2), 327–31.

[Bruce & Voi, 1983] Bruce, V. & Voi, M. E. L. (1983). Recognizing Faces [and Discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 302(1110), 423–436.

[Bruner, 1975] Bruner, J. (1975). The ontogenesis of speech acts. *Journal of child language*, 2(1), 1–19.

[Bugmann et al.,] Bugmann, G., Lauria, S., Kyriacou, T., EwanKlein, Bos, J., & Coventry, K. Using verbal instructions for route learning: Instruction analysis. In *TIMR 01, Towards Intelligent Mobile Robots*.

[Burke et al., 2002] Burke, C., Harper, L., & Loehr, D. (2002). A flexible architecture for a multimodal robot control interface. In *AAAI Technical Report WS-02-08*.

[Busso et al., 2004] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., & Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces - ICMI '04* (pp. 205). New York, New York, USA: ACM Press.

[C. Breazeal, 2000] C. Breazeal, B. S. (2000). Infant-like social interactions between a robot and a human caregiver. *Adaptive Behavior*, (pp. 49–74).

[Callejas & Lopezcozar, 2008] Callejas, Z. & Lopezcozar, R. (2008). Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication*, 50(5), 416–433.

[Carnegie & Kiesler, 2002] Carnegie, S. K. & Kiesler, S. (2002). Mental models and cooperation with robotic assistants. In *In Proceedings of Conference on Human Factors in Computing Systems* (pp. 576–577).: ACM Press.

[Cassimatis et al., 2004] Cassimatis, N., Trafton, J., Bugajska, M., & Schultz, A. (2004). *Integrating Cognition, Perception and Action through Mental Simulation in Robots*. Technical report, Naval Research Laboratory.

[Castro-González.,] Castro-González., A. *Bio-inspired Decision Making System for an Autonomous Social Robot. The role of fear.* PhD thesis, Carlos III University of Madrid.

[Chauhan & Lopes, 2010] Chauhan, A. & Lopes, L. S. (2010). Acquiring vocabulary through human robot interaction: A learning architecture for grounding words with multiple meanings. In *AAAI Fall Symposium: Dialog with Robots* Arlington, VA, USA: AAAI Press AAAI Press.

[Chen et al., 1998] Chen, L., Huang, T., Miyasato, T., & Nakatsu, R. (1998). Multimodal human emotion/expression recognition. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 366–371).: IEEE Comput. Soc.

[Chetty, 2009] Chetty, V. (2009). Microsoft's Project Natal for Xbox 360.

[Cheyer et al., 1998] Cheyer, A., Julia, L., Bunt, H., Beun, R.-J., & Borghuis, T. (1998). *Multimodal Human-Computer Communication*, volume 1374 of *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg.

[Clark, 1996] Clark, H. (1996). *Uses of Language*. Cambridge: Cambridge University Press.

[Conway et al., 2001] Conway, A., Cowan, N., & Bunting, M. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*.

[Cowie & Douglas-Cowie, 2000] Cowie, R. & Douglas-Cowie, E. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. In *In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

[Cowie et al., 2001] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1), 32–80.

[Cowie, R., Douglas-Cowie, E., & Romano et al., 1999] Cowie, R., Douglas-Cowie, E., & Romano, A. . C. e. t. i. d., its prosodic correlates. In ESCA Tutorial, on Dialogue, R. W. E., & Prosody. (1999). Changing emotional tone in dialogue and its prosodic correlates. *ESCA Tutorial and Research Workshop (ETRW) on Dialogue and Prosody*.

[Dalmasso et al.,] Dalmasso, E., Castaldo, F., Laface, P., Colibro, D., & Vair, C. Loquendo - Speaker recognition evaluation system. In *Acoustics, Speech and Signal Processing, ICASSP 2009. IEEE International Conference on* (pp. 4213–4216). Taipei.

[Davis, 1971] Davis, F. (1971). *Inside Intuition-What we know about Non-Verbal Communication*. McGraw-Hill Book Co.

[De Silva et al., 2007] De Silva, L., Miyasato, T., & Nakatsu, R. (2007). Facial emotion recognition using multi-modal information. In *Proceedings of ICICS,*, volume 1 (pp. 397–401).: IEEE.

[Dominey et al., 2007] Dominey, P. F., Mallet, A., & Yoshida, E. (2007). Progress in programming the hrp-2 humanoid using spoken language. In *International Conference on Robotics and Automation (ICRA07)*.

[Dooling & Popper, 2000] Dooling, R. & Popper, A. (2000). Hearing in birds and reptiles: an overview. *Comparative Hearing: Reptiles and Birds. Springer-Verlag, New York*, (pp. 1–12).

[Eberman et al., 2002] Eberman, B., Carter, J., Meyer, D., & Goddeau, D. (2002). Building voiceXML browsers with openVXI. In *Proceedings of the eleventh international conference on World Wide Web - WWW '02* (pp. 713). New York, New York, USA: ACM Press.

[Ekman & Friesen, 1971] Ekman, P. & Friesen, W. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology, 17*.

[Ekman et al., 1972] Ekman, P., Friesen, W., & Ellsworth, P. (1972). Emotion in the human face: Guidelines for research and an integration of findings.

[Ekman et al., 1978] Ekman, P., Friesen, W., & Hager, J. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. In *Consulting Psychologists Press*, number A Human Face Palo Alto.

[Ekman & R.J, 1994] Ekman, P. & R.J, D. (1994). The Nature of Emotion: Fundamental Questions.

[Eyben et al., 2009] Eyben, F., Wollmer, M., & Schuller, B. (2009). OpenEAR — Introducing the munich open-source emotion and affect recognition toolkit. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (pp. 1–6).: IEEE.

[Eyben et al., 2010] Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile. In *Proceedings of the international conference on Multimedia - MM '10* (pp. 1459). New York, New York, USA: ACM Press.

[Falb et al., 2007] Falb, J., Popp, R., Rock, T., Jelinek, H., Arnautovic, E., & Kaindl, H. (2007). Fully-automatic generation of user interfaces for multiple devices from a high-level model based on communicative acts. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07). Jan.* (pp. 26–26). Hawaii (EEUU): IEEE.

[Falb et al., 2006] Falb, J., Rock, T., & Arnautovic, E. (2006). Using communicative acts in interaction design specifications for automated synthesis of user interfaces.

In *21st IEEE/ACM International Conference on Automated Software Engineering (ASE'06). Sept.* (pp. 261–264). Tokyo (Japan): IEEE.

[Feldman, 1999] Feldman, S. (1999). NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. *Online-Weston then Wilton-*.

[Fiebrink & Cook, 2010] Fiebrink, R. & Cook, P. (2010). The Wekinator: A System for Real-time, Interactive Machine Learning in Music.

[Fisher & Byrne, 1975] Fisher, J. & Byrne, D. (1975). Too close for comfort: Sex differences in response to invasions of personal space. *Journal of Personality and Social Psychology*, 32(1), 15.

[Freeman & Boyd, 1993] Freeman, D. & Boyd, I. (1993). Voice activity detection. *EP Patent 0,335,521*.

[Freeman & Boyd, 1994] Freeman, D. & Boyd, I. (1994). Voice activity detection. *US Patent 5,276,765*.

[Fry et al., 1998] Fry, J., Asoh, H., & Matsui, T. (1998). Natural dialogue with the Jijo-2 office robot. In *Intelligent Robots and Systems, 1998. Proceedings., 1998 IEEE/RSJ International Conference on* (pp. 1278–1283). Victoria (Canada).

[Gauvain, 2002] Gauvain, J. (2002). The LIMSI broadcast news transcription system. *Speech Communication*.

[Gibbon, Moore, 1997] Gibbon, Moore, W. (1997). *Handbook of standards and resources for spoken language systems*.

[Gockley et al., 2005] Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., & Schultz, A. (2005). Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1338–1343).: IEEE.

[Goddeau & Pineau,] Goddeau, D. & Pineau, J. *Fast reinforcement learning of dialog strategies*. IEEE.

[Gonzalez-Pacheco et al., 2011] Gonzalez-Pacheco, V., Ramey, A., Alonso-Martin, F., Castro-Gonzalez, A., & Salichs, M. A. (2011). Maggie: A Social Robot as a Gaming Platform. *International Journal of Social Robotics*, 3(4), 371–381.

[Gorostiza et al., 2006a] Gorostiza, J., Barber, R., Khamis, A., Malfaz, M., Pacheco, R., Rivas, R., Corrales, A., Delgado, E., & Salichs, M. (2006a). Multimodal Human-Robot Interaction Framework for a Personal Robot. In *ROMAN 2006 - The 15th*

IEEE International Symposium on Robot and Human Interactive Communication. Sept (pp. 39–44). Hatfield (UK): IEEE.

[Gorostiza et al., 2006b] Gorostiza, J., Barber, R., Khamis, A., Malfaz, M., Pacheco, R., Rivas, R., Corrales, A., Delgado, E., & Salichs, M. (2006b). *Multimodal Human-Robot Interaction Framework for a Personal Robot.* IEEE.

[Gorostiza & Salichs, 2011] Gorostiza, J. F. & Salichs, M. A. (2011). End-user programming of a social robot by dialog. *Robotics and Autonomous Systems*, (59), 1102–1114.

[Graf et al., 2004] Graf, B., Hans, M., & Schraft, R. D. (2004). Care-O-bot II — Development of a Next Generation Robotic Home Assistant. *Autonomous Robots*, 16, 193–205.

[Grammer, 1998] Grammer, K. (1998). The courtship dance: Patterns of nonverbal synchronization in opposite-sex encounters. *Journal of Nonverbal Behavior.*

[Grosz & Sidner, 1986] Grosz, B. J. & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 175–204.

[Haasch et al., 2004] Haasch, A., Hohenner, S., Hüwel, S., Kleinehagenbrock, M., Lang, S., Toptsis, I., Fink, G. A., J. Fritsch, B. W., & Sagerer, G. (2004). Biron - the bielefeld robot companion. In *Proceedings International Workshop on Advances in Service Robotics* (pp. 27–32).

[Hall, 1966] Hall, E. (1966). The hidden dimension. *Garden City, N.Y*, Doubleday.

[Handzel & Krishnaprasad, 2002] Handzel, A. & Krishnaprasad, P. (2002). Biomimetic sound-source localization. *IEEE Sensors Journal*, 2(6), 607–616.

[Hayduk, 1978] Hayduk, L. (1978). Personal space: An evaluative and orienting overview. *Psychological Bulletin*, 85(1), 117.

[Haykin & Chen, 2005] Haykin, S. & Chen, Z. (2005). The cocktail party problem. *Neural computation.*

[Henderson et al., 2008] Henderson, J., Lemon, O., & Georgila, K. (2008). Hybrid reinforcement/supervides learning of dialogue policies from fixed data sets. *Computational Linguistics*, 34(4).

[Henkel, 2012] Henkel, Z. (2012). Towards A Computational Method of Scaling A Robot's Behavior via Proxemics. In *HRI '12 Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction. March* (pp. 145–146). Boston (EEUU).

[Holmes et al., 1994] Holmes, G., Donkin, A., & Witten, I. (1994). WEKA: a machine learning workbench. In *Proceedings of ANZIIS '94 - Australian New Zealnd Intelligent Information Systems Conference* (pp. 357–361).: IEEE.

[Hudspeth, 1983] Hudspeth, A. J. (1983). The hair cells of the inner ear. *Scientific American*, 248, 54–64.

[Hüttenrauch, 2006] Hüttenrauch, H. (2006). Investigating spatial relationships in human-robot interaction. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference. Oct* (pp. 5052 – 5059). Beijing (China).

[Hüwel et al., 2006] Hüwel, S., Wrede, B., & Sagerer, G. (2006). Robust Speech Understanding for Multi-Modal Human-Robot Communication. In *15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06)* (pp. 45–50).

[Iba et al., 2002] Iba, S., Paredis, C. J., & Khosla, P. K. (2002). Interactive multimodal robot programming. In *2002 IEEE International Conference on Robotics and Automation*: IEEE/RSJ.

[Ishi et al., 2006a] Ishi, C., Matsuda, S., & Kanda, T. (2006a). Robust speech recognition system for communication robots in real environments. *Humanoid Robots, ...*

[Ishi et al., 2006b] Ishi, C., Matsuda, S., Kanda, T., Jitsuhiro, T., Ishiguro, H., Nakamura, S., & Hagita, N. (2006b). Robust Speech Recognition System for Communication Robots in Real Environments. In *2006 6th IEEE-RAS International Conference on Humanoid Robots* (pp. 340–345).: IEEE.

[Ishiguro et al., 2002] Ishiguro, H., Miyashita, T., Kanda, T., Ono, T., & Imai, M. (2002). Robovie: An interactive humanoid robot - toward new information media support communications. In *Video Proceedings of IEEE Int. Conf. Robotics and Automation (ICRA).*

[Iwahashi et al., 2010] Iwahashi, N., Sugiura, K., Taguchi, R., Nagai, T., & Taniguchi, T. (2010). Robots that learn to communicate: A developmental approach to personally and physically situated human-robot conversations. In *AAAI Fall Symposium: Dialog with Robots* Arlington, VA, USA: AAAI Press AAAI Press.

[Izard, 1990] Izard, C. (1990). Facial expressions and the regulation of emotions. *Journal of personality and social psychology, 58.*

[Jansen & Belpaeme, 2006] Jansen, B. & Belpaeme, T. (2006). A computational model of intention reading in imitation. *Robotics and Autonomous Systems*, 54(5), 394–402.

[J.F.Gorostiza., 2010] J.F.Gorostiza. (2010). *Programación natural de un robot social mediante diálogos*. PhD thesis, Universidad Carlos III Madrid.

[Jokinen, 2003] Jokinen, K. (2003). Natural interaction in spoken dialogue systems. In *Proceedings of the Workshop Ontologies and Multilinguality in User Interfaces. HCI International 2003* (pp. 730–734).: Citeseer.

[Jurafsky & Martin, 2000] Jurafsky, D. & Martin, J. H. (2000). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. New Jersey, USA: Prentice-Hall, Inc.

[Jwu-Sheng et al., 2009] Jwu-Sheng, H., Chen-Yu, C., Cheng-Kang, W., & Chieh-Chih, W. (2009). Simultaneous localization of mobile robot and multiple sound sources using microphone array. In *2009 IEEE International Conference on Robotics and Automation. May* (pp. 29–34). Kobe (Japan): IEEE.

[Kanda et al., 2010] Kanda, T., Shiomi, M., Miyashita, Z., Ishiguro, H., & Hagita, N. (2010). A communication robot in a shopping mall. *IEEE Transactions on Robotics*, 26(5).

[Kearney & McKenzie, 1993] Kearney, G. & McKenzie, S. (1993). Machine interpretation of emotion: Design of a memory-based expert system for interpreting facial expressions in terms of signaled emotions. *Cognitive Science*.

[Kendon, 1970] Kendon, A. (1970). Movement coordination in social interaction: Some examples described. *Acta psychologica*.

[Khalili & Moradi, 2009] Khalili, Z. & Moradi, M. (2009). Emotion recognition system using brain and peripheral signals: using correlation dimension to improve the results of EEG. *Neural Networks, 2009. IJCNN 2009. ....*

[Kheng Lee et al., 2007] Kheng Lee, K., Syrdal, D., Walters, M., & Dautenhahn, K. (2007). Living with Robots: Investigating the Habituation Effect in Participants' Preferences During a Longitudinal Human-Robot Interaction Study. In *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on. Aug* (pp. 564–569).

[Kibria & Hellström, 2007] Kibria, S. & Hellström, T. (2007). Voice user interface in robotics - common issues and problems. *aass.oru.se*.

[Kidd et al., 2004] Kidd, C., Lee, H., Lieberman, J., & Lockerd, A. (2004). Humanoid Robots as Cooperative Partners for People. *International Journal of Humanoid Robotics*, 1(2), 1–34.

[Kidd & Breazeal, 2008] Kidd, C. D. & Breazeal, C. (2008). Robots at Home : Understanding Long-Term Human-Robot Interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 3230–3235).

[Kim & Choi, 2007] Kim, H. & Choi, J. (2007). Human-robot interaction in real environments by audio-visual integration. *International Journal of Control*, 5(1), 61–69.

[Kobayashi & Hara, 1997] Kobayashi, H. & Hara, F. (1997). Facial interaction between animated 3D face robot and human beings. In *IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 4 (pp. 3732–3737).: IEEE.

[Kruijff et al., 2007] Kruijff, G.-J. M., Zender, H., Jensfelt, P., , & Christensen, H. I. (2007). Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems, Special Issue on Human and Robot Interactive Communication*, 4(2).

[Kruijff et al., 2006] Kruijff, G.-J. M., Zender, H., Jensfelt, P., & Christensen, H. I. (2006). Clarification dialogues in human-augmented mapping. In *HRI '06: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (pp. 282–289). New York, NY, USA: ACM.

[Küblbeck & Ernst, 2006] Küblbeck, C. & Ernst, A. (2006). Face detection and tracking in video sequences using the modifiedcensus transformation. *Image and Vision Computing*.

[Lambert,] Lambert, D. the Diagram Group.(1996). Body Language. *Glasgow: Harper Collins*.

[Lambert, 2004] Lambert, D. (2004). Body Language. *Harper Collins, London*.

[Lamel, 2002] Lamel, L. (2002). Lightly supervised and unsupervised acoustic model training. *Computer Speech & Language*.

[Larsson et al., 2004] Larsson, S., Berman, A., Bos, J., Grönqvist, L., Ljunglöf, P., & Traum, D. (2004). *TRINDIKIT 3.1 Manual*. Goteborg, Suiza: Goteborg University.

[Laurence & Nigay. Coutaz, 1993] Laurence & Nigay. Coutaz, J. (1993). A design space for multimodal systems: concurrent processing and data fusion. In ACM (Ed.), *Proceedings of the INTERACT'93 and CHI'93* (pp. 172–178).

[Lauria et al., 2001] Lauria, S., Bugmann, G., Kyriacou, T., Bos, J., & Klein, E. (2001). Training personal robots using natural language instruction. *IEEE Intelligent Systems*, 16, 38–45.

[Lauria et al., 2002] Lauria, S., Bugmann, G., Kyriacou, T., & Klein, E. (2002). Mobile robot programming using natural language. *Robotics and Autonomous Systems*, 38, 171–181.

[Lemon et al., 2002] Lemon, O., Gruenstein, A., Battle, A., & Peters, S. (2002). Multi-tasking and collaborative activities in dialogue systems. In *Proceedings of the 3rd SIGdial workshop on Discourse and dialogue - Volume 2*, SIGDIAL '02 (pp. 113–124). Stroudsburg, PA, USA: Association for Computational Linguistics.

[Liddy, 2005] Liddy, E. D. (2005). Enhanced Text Retrieval Using Natural Language Processing. *Bulletin of the American Society for Information Science and Technology*, 24(4), 14–16.

[Litman & Forbes-Riley, 2004] Litman, D. J. & Forbes-Riley, K. (2004). Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04* (pp. 351–es). Morristown, NJ, USA: Association for Computational Linguistics.

[Littlewort et al., 2011] Littlewort, G., Whitehill, J., Wu, T.-F., Butko, N., Ruvolo, P., Movellan, J., & Bartlett, M. (2011). The motion in emotion — A CERT based approach to the FERA emotion challenge. In *Face and Gesture 2011* (pp. 897–902).: IEEE.

[Llisterri et al., 2003] Llisterri, J., Carbó, C., Machuca, M., De la Mota, C., Riera, M., & R\'\ios, A. (2003). El papel de la fonética en el desarrollo de las tecnologías del habla. In *Memorias de las VII Jornadas de Linguística.* Cadiz (Spain): Servicio de Publicaciones de la Universidad de Cádiz.

[Looije et al., 2010] Looije, R., Neerincx, M. a., & Cnossen, F. (2010). Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *International Journal of Human-Computer Studies*, 68(6), 386–397.

[López-Cózar & Callejas, 2006] López-Cózar, R. & Callejas, Z. (2006). DS-UCAT: Sistema de diálogo multimodal y multilingüe para un entorno educativo. *Proc. IV Jornadas en ....*

[Lucas, 2000] Lucas, B. (2000). VoiceXML for Web-based distributed conversational applications. *Communications of the Association for Computing Machinery (ACM)*, 43(9), 53–57.

[Lucas Cuesta et al., 2008] Lucas Cuesta, J., Alcázar Prior, R., Montero Martínez, J., Fernández Martínez, F., Barra Chicote, R., D'Haro Enriquez, L., Ferreiros López, J., Córdoba Herralde, R., Macias Guarasa, J., San Segundo Hernández, R., & Others (2008). Desarrollo de un robot-gu\'\ia con integración de un sistema de diálogo y expresión de emociones: Proyecto robint. *Procesamiento del lenguaje natural*, (40), 51–58.

[Lucey et al., 2006] Lucey, S., Matthews, I., & Hu, C. (2006). AAM derived face representations for robust facial action recognition. In *In Automatic Face and Gesture Recognition, 2006. 7th International Conference on*.

[Malfaz & Salichs, 2004] Malfaz, M. & Salichs, M. A. (2004). : Lisboa, Portugal: Fifth IFAC Symposium on Intelligent Autonomous Vehicles.

[Marge et al., 2009] Marge, M., Pappu, A., Frisch, B., Harris, T., & Rudnicky, A. (2009). Exploring Spoken Dialog Interaction in Human-Robot Teams.

[Martinovsky, 2006] Martinovsky, B. (2006). *The error is the clue: Breakdown in human-machine interaction*. Technical report, DTIC Document.

[Minami et al., 2010] Minami, Y., Higashinaka, R., Dohsaka, K., Meguro, T., & Maeda, E. (2010). Trigram dialogue control using POMDPs. In *2010 IEEE Spoken Language Technology Workshop* (pp. 336–341).: IEEE.

[Mitsunaga et al., 2006] Mitsunaga, N., Miyashita, T., Ishiguro, H., Kogure, K., & Hagita, N. (2006). Robovie-IV: A Communication Robot Interacting with People Daily in an Office. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on* (pp. 5066–5072). Beijing: IEEE.

[Morris, 1946] Morris, C. (1946). *Signs, Language and Behavior*. New York, USA: Prentice-Hall.

[Morrongiello, 1989] Morrongiello, B. A. (1989). Infants' monaural localization of sounds: effects of unilateral ear infection. *The Journal of the Acoustical Society of America*, 86(2), 597–602.

[Mumm & Mutlu, 2011] Mumm, J. & Mutlu, B. (2011). Human-robot proxemics. In *Proceedings of the 6th international conference on Human-robot interaction - HRI '11. March* (pp. 331). Lausanne (Switzerland): ACM Press.

[Murray et al., 2004] Murray, J., Erwin, H., & Wermter, S. (2004). Robotics sound-source localization and tracking using interaural time difference and crosscorrelation. In *Proceedings of NeuroBotics Workshop. Jun.* (pp. 89–97). Ulm (Germany).

[Murugappan & Rizon, 2008] Murugappan, M. & Rizon, M. (2008). Time-frequency analysis of EEG signals for human emotion detection. *In 4th Kuala Lumpur International Conference on Biomedical Engineering*, (pp. 262–265).

[Nakadai et al.,] Nakadai, K., Matsuura, D., Okuno, H., & Kitano, H. Applying scattering theory to robot audition system: Robust sound source localization and extraction. In *Intelligent Robots and Systems, 2003.(IROS 2003). Oct. 2003*, volume 2 (pp. 1147–1152). Las Vegas (EEUU): IEEE.

[Nakadai et al., 2002] Nakadai, K., Okuno, H., & Kitano, H. (2002). Real-time sound source localization and separation for robot audition. In *Seventh International Conference on Spoken Language Processing* (pp. 193–196).

[Nakadai et al., 2008a] Nakadai, K., Okuno, H. G., Nakajima, H., Hasegawa, Y., & Tsujino, H. (2008a). An Open Source Software System For Robot Audition HARK and Its Evaluation. In *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on. Dec* Daejeon (Korea).

[Nakadai et al., 2008b] Nakadai, K., Okuno, H. G., Nakajima, H., Hasegawa, Y., & Tsujino, H. (2008b). An open source software system for robot audition HARK and its evaluation. In *Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots* (pp. 561–566).: IEEE.

[Nass et al., 1994] Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *CHI '94 Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence. April* (pp. 72–78). New York, NY, USA.

[Nass, C; Reeves, 1996] Nass, C; Reeves, B. (1996). *How people treat computers, television, and new media like real people and places.*

[Newcomb, 1953] Newcomb, T. (1953). An approach to the study of communicative acts. *Psychological review.*

[Niklfeld et al., 2001] Niklfeld, G., Finan, R., & Pucher, M. (2001). Architecture for adaptive multimodal dialog systems based on voiceXML. *INTERSPEECH*, (1), 2341–2344.

[Nyberg et al., 2002] Nyberg, E., Mitamura, T., Placeway, P., Duggan, M., & Francisco, S. (2002). DialogXML: Extending VoiceXML for Dynamic Dialog Management. In *Proceedings of the second international conference on Human Language Technology Research* (pp. 298–302). San Diego (California): Morgan Kaufmann Publishers Inc.

[Oh et al., 1992] Oh, S., Viswanathan, V., & Papamichalis, P. (1992). Hands-free voice communication in an automobile with a microphone array. In *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 281–284). San Francisco: IEEE.

[Padgett & Cottrell, 1997] Padgett, C. & Cottrell, G. (1997). Representing face images for emotion classification. *Advances in neural information processing systems*.

[Pantic et al., 2005] Pantic, M., Sebe, N., Cohn, J. F., & Huang, T. (2005). Affective multimodal human-computer interaction. In *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05* (pp. 669). New York, New York, USA: ACM Press.

[Pantie & Rothkrantz, 2000] Pantie, M. & Rothkrantz, L. (2000). Automatic analysis of facial expressions: the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12), 1424–1445.

[Paolo Baggia, 2005] Paolo Baggia, S. M. (2005). Loquendo Speech Technologies and multimodality.

[Paul Watzlawick, 1967] Paul Watzlawick, Janet Helmick Beavin, D. D. J. (1967). *Pragmatics of Human Communication. A study of Interactional patterns, pathologies and paradoxes*. W. W. Norton and Co.

[Pelachaud et al., 1994] Pelachaud, C., Badler, N., & Viaud, M. (1994). Final report to NSF of the standards for facial animation workshop.

[Peltason & Wrede, 2010a] Peltason, J. & Wrede, B. (2010a). Modeling human-robot interaction based on generic interaction patterns. In *AAAI Fall Symposium: Dialog with Robots* Arlington, VA, USA: AAAI Press AAAI Press.

[Peltason & Wrede, 2010b] Peltason, J. & Wrede, B. (2010b). Pamini: A framework for assembling mixed-initiative human-robot interaction from generic interaction patterns. In *SIGDIAL Conference* (pp. 229–232).

[Perzanowski et al., 2001a] Perzanowski, D., Schultz, A., Adams, W., Marsh, E., & Bugajska, M. (2001a). Building a multimodal human-robot interface. *IEEE Intelligent Systems*, 16(1), 16–21.

[Perzanowski et al., 2001b] Perzanowski, D., Schultz, A. C., Adams, W., Marsh, E., & Bugajska, M. (2001b). Building a multimodal human–robot interface.

[Peto, 1980] Peto, J. (1980). Manyears: a program for computing observed and expected deaths or incidence rates. *International publication. Oxford: ICRF Cancer Unit*.

[Petrantonakis, Panagiotis C. & Hadjileontiadis., 2010] Petrantonakis, Panagiotis C. & Hadjileontiadis., L. J. (2010). Emotion recognition from EEG using higher order crossings. *Information Technology in Biomedicine, IEEE Transactions on 14.2*, (pp. 186–197).

[Picard, 2000] Picard, R. (2000). *Affective computing*. MIT Press.

[Pineau, 2003] Pineau, J. (2003). Towards robotic assistants in nursing homes: Challenges and results. *Robotics and Autonomous Systems*, 42(3-4), 271–281.

[Plutchik, 1980] Plutchik (1980). Emotion, a psychoevolutionary synthesis.

[Qi, 2008] Qi, Z. (2008). *Real-time adaptive noise cancellation for automatic speech recognition in a car environment*. PhD thesis.

[Quigley et al., 2009] Quigley, M., Gerkey, B., Conley, K., Faust, J., Foote, T., Leibs, J., Berger, E., Wheeler, R., & Ng, A. (2009). ROS : an open-source Robot Operating System. In *ICRA workshop on open source software*, number Figure 1 (pp.5̃). Kobe (Japan).

[R. Rivas, A. Corrales, R. Barber, 2007] R. Rivas, A. Corrales, R. Barber, M. A. S. (2007). Robot Skill Abstraction for AD Architecture.

[Ramírez, 2009] Ramírez, A. P. (2009). Uso del Cry Translator (Traductor del llanto del bebé) de Biloop Technologic SL (España) como identificador del llanto en el niño y pautas a seguir. *Clínica*.

[Reithinger & Alexandersson, 2003] Reithinger, N. & Alexandersson, J. (2003). SmartKom: adaptive and flexible multimodal access to multiple applications. *ICMI '03 Proceedings of the 5th international conference on Multimodal interfaces*.

[Rich & Ponsler, 2010] Rich, C. & Ponsler, B. (2010). Recognizing engagement in human-robot interaction. In *In Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on* (pp. 375–382).

[Ross & Smith, 1978] Ross, R. J. & Smith, J. J. B. (1978). Detection of substrate vibrations by salamanders: inner ear sense organ activity. *Canadian Journal of Zoology*, 56(5), 1156–1162.

[Roy & Pentland, 1996] Roy, D. & Pentland, A. (1996). Automatic spoken affect classification and analysis. In *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition* (pp. 363–367).: IEEE Comput. Soc. Press.

[Roy et al., 1998] Roy, N., Pineau, J., & Thrun, S. (1998). Spoken Dialog Management for Robots. *Management*.

[Roy et al., 2000] Roy, N., Pineau, J., & Thrun, S. (2000). *Spoken dialogue management using probabilistic reasoning*. Morristown, NJ, USA: Association for Computational Linguistics.

[Rudnicky & Thayer, 1999] Rudnicky, A. & Thayer, E. (1999). Creating natural dialogs in the Carnegie Mellon Communicator system. *Sixth European ...*.

[Russell & Dols, 1997] Russell, J. & Dols, J. (1997). The psychology of facial expression.

[Sakagami et al., 2002] Sakagami, Y., Watanabe, R., Aoyama, C., Matsunaga, S., Higaki, N., & Fujimura, K. (2002). The intelligent ASIMO: system overview and integration. In *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on* (pp. 2478–2483). Laussane: IEEE.

[Salichs et al., 2006a] Salichs, M., Barber, R., Khamis, A., Malfaz, M., Gorostiza, J., Pacheco, R., Rivas, R., Corrales, A., Delgado, E., & Garcia, D. (2006a). *Maggie: A Robotic Platform for Human-Robot Social Interaction*. IEEE.

[Salichs et al., 2006b] Salichs, M. A., Barber, R., Khamis, A., Malfaz, M., Gorostiza, J., Pacheco, R., Rivas, R., Corrales, A., Delgado, E., & García, D. (2006b). Maggie: A robotic platform for human-robot social interaction. In *Submitted to IEEE International Conference on Robotics, Automation and Mechatronics (RAM 2006)* Bankgok, Thailand: IEEE.

[Saxena & Ng, 2009] Saxena, A. & Ng, A. (2009). Learning sound location from a single microphone. In *2009 IEEE International Conference on Robotics and Automation. May* (pp. 1737–1742). Kobe (Japan): IEEE.

[Schegloff & Sacks, 1973] Schegloff, E. A. & Sacks, H. (1973). Opening up closings. *Semiotica*, 8, 289 – 327.

[Schmitz et al., 2008] Schmitz, N., Hirth, J., & Berns, K. (2008). Realization of natural interaction dialogs in public environments using the humanoid robot ROMAN. In *Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots* (pp. 579–584).: IEEE.

[Schramm, 1954] Schramm, W. (1954). How communication works. *Process and Effects of Communication*, (pp. 3–26).

[Schuller & Arsic, 2006] Schuller, B. & Arsic, D. (2006). Emotion recognition in the noise applying large acoustic feature sets. *Speech Prosody, Dresden.*

[Schuller et al., 2004] Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *In Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on.*

[Schultz, 2004] Schultz, A. (2004). Cognitive tools for humanoid robots in space. In *Proceedings of the 16th IFAC Conference on Automatic Control in Aerospace* Elsevier, Oxford.

[Searle, 1969] Searle, J. (1969). *Speech acts: An essay in the philosophy of language.* Cambridge Univ Pr.

[Searle, 1975] Searle, J. (1975). Indirect speech acts. *Syntax and semantics*, 3, 59–82.

[Sebe et al., 2006] Sebe, N., Cohen, I., Gevers, T., & Huang, T. (2006). Emotion Recognition Based on Joint Visual and Audio Cues. In *18th International Conference on Pattern Recognition (ICPR'06)* (pp. 1136–1139).: IEEE.

[Seneff et al., 1996] Seneff, S., Goddeau, D., Pao, C., & Polifroni, J. (1996). Multimodal discourse modelling in a multi-user multi-domain environment. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 1 (pp. 192–195).: IEEE.

[Shaffer, 1982] Shaffer, L. (1982). Rhythm and timing in skill. *Psychological Review; Psychological Review.*

[Shannon, 1948] Shannon, C. E. (1948). A mathematic theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656.

[Shimokawa & Sawaragi, 2001] Shimokawa, T. & Sawaragi, T. (2001). Acquiring communicative motor acts of social robot using interactive evolutionary computation. In *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236)*, volume 3 (pp. 1396–1401).: IEEE.

[Shuyin et al., 2004] Shuyin, I. T., Toptsis, I., Li, S., Wrede, B., & Fink, G. A. (2004). A Multi-modal Dialog System for a Mobile Robot. In *Int. Conf. on Spoken Language Processing* (pp. 273–276). Jeju Island (Korea): IEEE.

[Skubic et al., 2004] Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., Bugajska, M., & Brock, D. (2004). Spatial language for human–robot dialogs. *IEEE Transactions on Systems, Man and Cybernetics. Part-C: Applications and Reviews*, 34.

[Spexard et al., 2006] Spexard, T., Li, S., Wrede, B., Fritsch, J., Sagerer, G., Booij, O., Zivkovic, Z., Terwijn, B., & Krose, B. (2006). BIRON, where are you? Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 934–940).: IEEE.

[Steidl, 2009] Steidl, S. (2009). *Automatic Classification of Emotion Related User States in Spontaneous Children's Speech*. Logos-Verlag.

[Stiefelhagen, 2004] Stiefelhagen, R. (2004). Natural human-robot interaction using speech, head pose and gestures. In IEEE (Ed.), *Intelligent Robots and Systems, 2004. (IROS 2004).*, volume 3 (pp. 2422– 2427 vol.3).

[Taguchi et al., 2009] Taguchi, R., Iwahashi, N., & Nitta, T. (2009). : chapter Learning Communicative Meanings of Utterances by Robots, (pp. 62–72).

[Takahashi et al., 2010] Takahashi, T., Nakadai, K., Komatani, K., Ogata, T., & Okuno, H. (2010). Improvement in listening capability for humanoid robot HRP-2. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on. May* (pp. 470–475). Anchorage (EEUU): IEEE.

[Takayama & Pantofaru, 2009] Takayama, L. & Pantofaru, C. (2009). Influences on proxemic behaviors in human-robot interaction. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems. Oct* (pp. 5495–5502). Menlo Park, CA, USA: IEEE.

[Tamai et al., 2005] Tamai, Y., Sasaki, Y., Kagami, S., & Mizoguchi, H. (2005). Three Ring Microphone Array for 3D Sound Localization and Separation for Mobile Robot Audition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 903–908). Edmonton (Canada): IEEE.

[Tanaka et al., 2010] Tanaka, N., Ogawa, T., Akagiri, K., & Kobayashi, T. (2010). DEVELOPMENT OF ZONAL BEAMFORMER AND ITS APPLICATION TO ROBOT AUDITION. In *Signal Processing*, volume 1 (pp. 1529–1533).

[Tellex et al., 2011] Tellex, S., Kollar, T., Dickerson, S., Walter, M. R., Banerjee, A. G., Teller, S., & Roy, N. (2011). Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine*, 32(4), 64–76.

[Terzopoulos & Waters, 1993] Terzopoulos, D. & Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

[Tomasello, 2008] Tomasello, M. (2008). *Origins of Human Communication*. Cambridge, Massachusetts: The MIT Press.

[Tong et al., 2007] Tong, Y., Liao, W., & Ji, Q. (2007). Facial action unit recognition by exploiting their dynamic and semantic relationships. *attern Analysis and Machine Intelligence, IEEE Transactions on*.

[Toptsis et al., 2004] Toptsis, I., Li, S., Wrede, B., & Fink, G. A. (2004). A multimodal dialog system for a mobile robot. In *International Conference on Spoken Language Processing*, volume 1 (pp. 273–276).

[Trafton et al., 2006] Trafton, J. G., Schultz, A. C., Perznowski, D., Bugajska, M. D., Adams, W., Cassimatis, N. L., & Brock, D. P. (2006). Children and robots learning to play hide and seek. In *HRI '06: Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (pp. 242–249). New York, NY, USA: ACM Press.

[Traum, 2000] Traum, S. L. D. (2000). Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural Language Engineering*, 6, 323 – 340.

[Truong et al., 2007] Truong, K., van Leeuwen, D., & Neerincx, M. (2007). Unobtrusive multimodal emotion detection in adaptive interfaces: speech and facial expressions. *In Foundations of Augmented Cognition*, Springer, 354–363).

[Tu & Yu, 2012] Tu, B. & Yu, F. (2012). Bimodal Emotion Recognition Based on Speech Signals and Facial Expression. *Foundations of Intelligent Systems*.

[Turunen et al., 2005] Turunen, M., Hakulinen, J., Raiha, K.-J., Salonen, E.-P., Kainulainen, A., & Prusi, P. (2005). An architecture and applications for speech-based accessibility systems. *IBM Systems Journal*, 44(3), 485–504.

[Valderrama Cuadros, C. E., Ulloa Villegas, 2012] Valderrama Cuadros, C. E., Ulloa Villegas, G. V. (2012). Spectral analysis of physiological parameters for consumers' emotion detection.

[Valin et al., 2004] Valin, J., Michaud, F., Hadjou, B., & Rouat, J. (2004). Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach. In *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, volume 1 (pp. 1033–1038).: IEEE.

[Valin et al., 2007] Valin, J., Michaud, F., & Rouat, J. (2007). Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55(3), 216–228.

[Valin et al., a] Valin, J., Michaud, F., Rouat, J., & Létourneau, D. Robust sound source localization using a microphone array on a mobile robot. In *Intelligent Robots and Systems, 2003.(IROS 2003). Oct. 2003*, volume 2 (pp. 1228–1233). Las Vegas (EEUU): IEEE.

[Valin et al., b] Valin, J.-M., Rouat, J., & Michaud, F. *Enhanced robot audition based on microphone array source separation with post-filter*. IEEE.

[Viola & Jones, 2004] Viola, P. & Jones, M. (2004). Robust real-time face detection. *International journal of computer vision*.

[Vlasenko & Schuller, 2007a] Vlasenko, B. & Schuller, B. (2007a). Combining frame and turn-level information for robust recognition of emotions within speech. *In Proceedings of Interspeech*.

[Vlasenko & Schuller, 2007b] Vlasenko, B. & Schuller, B. (2007b). Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. *In Affective Computing and Intelligent Interaction*.

[Wahlster, 2001a] Wahlster, W. (2001a). SmartKom: Multimodal communication with a life-like character. *Proceedings of the 7th European Conference on Speech Communication and Technology*.

[Wahlster, 2001b] Wahlster, W. (2001b). Smartkom: Towards multimodal dialogues with anthropomorphic interface agents. *Proceedings of the International Status Conference "Human-Computer Interaction"*.

[Wahlster, 2003a] Wahlster, W. (2003a). Smartkom: Symmetric multimodality in an adaptive and reusable dialogue shell. *Proceedings of the Human Computer Interaction Status*.

[Wahlster, 2003b] Wahlster, W. (2003b). Towards symmetric multimodality: Fusion and fission of speech, gesture, and facial expression. *KI 2003: Advances in Artificial Intelligence*, (pp. 1–18).

[Wahlster, 2006] Wahlster, W. (2006). SmartKom: Foundations of Multimodal Dialogue Systems (Cognitive Technologies).

[Waibel & Suhm, 1997] Waibel, A. & Suhm, B. (1997). Multimodal interfaces for multimedia information agents. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1 (pp. 167–170).: IEEE Comput. Soc. Press.

[Wainer et al.,] Wainer, J., Ferrari, E., Dautenhahn, K., & Robinsn, B. The effectiveness of using a robotics class to foster collaboration among groups of children with autism in an exploratory study. *Journal of Personal and Ubiquitous Computing*, 14.

[Walker et al., 2001] Walker, M. A., Passonneau, R., & Boland, J. E. (2001). Quantitative and qualitative evaluation of Darpa Communicator spoken dialogue systems. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01* (pp. 515–522). Morristown, NJ, USA: Association for Computational Linguistics.

[Wallace, 2000] Wallace, R. (2000). Artificial linguistic internet computer entity (a.l.i.c.e.).

[Walters, 2009] Walters, M. (2009). An empirical framework for human-robot proxemics. *Computer Science UH Home collection for Pure.*

[Weizenbaum, 1966] Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1), 36–45.

[Wierzbicki et al., 2012] Wierzbicki, R., Tschoeppe, C., Ruf, T., & Garbas, J. (2012). EDIS-Emotion-Driven Interactive Systems. In *Semantic Ambient Media Workshop in conjunction with Pervasive* Newcastle, UK.

[Williams, 1971] Williams, J. (1971). Personal space and its relation to extraversion-introversion. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, (pp. 156–160).

[Williams & Young, 2007] Williams, J. D. & Young, S. (2007). Scaling POMDPs for Spoken Dialog Management. *IEEE Transactions on Audio, Speech and Language Processing*, 15(7), 2116–2129.

[Wood & Cowan, 1995] Wood, N. & Cowan, N. (1995). The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology-Learning Memory and Cognition, 21.*

[Wright & Fitzgerald, 2001] Wright, B. A. & Fitzgerald, M. B. (2001). Different patterns of human discrimination learning for two interaural cues to sound-source location. *Proceedings of the National Academy of Sciences of the United States of America*, 98(21), 12307–12.

[Yager & Hoy, 1986] Yager, D. & Hoy, R. (1986). The cyclopean ear: a new sense for the praying mantis. *Science*, 231(4739), 727–729.

[Yoshida et al., 2009] Yoshida, T., Nakadai, K., & Okuno, H. G. (2009). Automatic speech recognition improved by two-layered audio-visual integration for robot audition. In *2009 9th IEEE-RAS International Conference on Humanoid Robots* (pp. 604–609). Paris (France): IEEE.

[Yoshitomi et al., 2000] Yoshitomi, Y., Kawano, T., & Kilazoe, T. (2000). Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In *Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No.00TH8499)* (pp. 178–183).: IEEE.

[Young, 2006] Young, S. (2006). Using POMDPS for Dialog Management. In *2006 IEEE Spoken Language Technology Workshop* (pp. 8–13). Palm Beach (Aruba): IEEE.

[Young et al., 2010] Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., & Yu, K. (2010). The Hidden Information State model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2), 150–174.

[Zaslavsky, 2003] Zaslavsky, A. (2003). Communicative acts of Elvin-enhanced mobile agents. In *IEEE/WIC International Conference on Intelligent Agent Technology, 2003. IAT 2003.* (pp. 446–449).: IEEE Comput. Soc.

[Zhou & Yuan, 2010] Zhou, W. & Yuan, B. (2010). *Trainbot: a spoken dialog system using partially observable Markov decision processes*. IET.